SIAM J. OPTIM.
Vol. 18, No. 1, pp. 1–28

APPROXIMATION BOUNDS FOR QUADRATIC OPTIMIZATION WITH HOMOGENEOUS QUADRATIC CONSTRAINTS*

ZHI-QUAN LUO†, NICHOLAS D. SIDIROPOULOS‡, PAUL TSENG§, AND SHUZHONG ZHANG¶

Abstract. We consider the NP-hard problem of finding a minimum norm vector in n-dimensional real or complex Euclidean space, subject to m concave homogeneous quadratic constraints. We show that a semidefinite programming (SDP) relaxation for this nonconvex quadratically constrained quadratic program (QP) provides an $O(m^2)$ approximation in the real case and an $O(m)$ approximation in the complex case. Moreover, we show that these bounds are tight up to a constant factor. When the Hessian of each constraint function is of rank 1 (namely, outer products of some given so-called steering vectors) and the phase spread of the entries of these steering vectors are bounded away from $\pi/2$, we establish a certain "constant factor" approximation (depending on the phase spread but independent of m and n) for both the SDP relaxation and a convex QP restriction of the original NP-hard problem. Finally, we consider a related problem of finding a maximum norm vector subject to m convex homogeneous quadratic constraints. We show that an SDP relaxation for this nonconvex QP provides an $O(1/\ln(m))$ approximation, which is analogous to a result of Nemirovski et al. [*Math. Program.*, 86 (1999), pp. 463–473] for the real case.

Key words. semidefinite programming relaxation, nonconvex quadratic optimization, approximation bound

AMS subject classifications. 90C22, 90C20, 90C59

DOI. 10.1137/050642691

1. Introduction.

Consider the quadratic optimization problem with concave homogeneous quadratic constraints:

$$\begin{array}{rll} v_{\rm qp} := \min & \|z\|^2 & \\ \text{subject to (s.t.)} & \displaystyle\sum_{\ell\in\mathcal{I}_i} |h_\ell^H z|^2 \ge 1, & i=1,\dots,m, \\ & z\in\mathbb{F}^n, & \end{array} \tag{1}$$

where $\mathbb{F}$ is either $\mathbb{R}$ or $\mathbb{C}$, $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{F}^n$, $m \ge 1$, each h_ℓ is a given vector in $\mathbb{F}^n$, and $\mathcal{I}_1,\dots,\mathcal{I}_m$ are nonempty, mutually disjoint index sets satisfying $\mathcal{I}_1 \cup \cdots \cup \mathcal{I}_m = \{1,\dots,M\}$. Throughout, the superscript H will denote the complex Hermitian transpose, i.e., for $z = x + \mathbf{i}y$, where $x, y \in \mathbb{R}^n$ and $\mathbf{i}^2 = -1$, $z^H = x^T - \mathbf{i}y^T$. Geometrically, problem (1) corresponds to finding a least norm vector

*Received by the editors October 14, 2005; accepted for publication (in revised form) July 6, 2006; published electronically February 2, 2007.

http://www.siam.org/journals/siopt/18-1/64269.html

†Department of Electrical and Computer Engineering, University of Minnesota, 200 Union Street SE, Minneapolis, MN 55455 (luozq@ece.umn.edu). This author was supported in part by National Science Foundation grant DMS-0312416.

‡Department of Electronic and Computer Engineering, Technical University of Crete, 73100 Chania, Crete, Greece (nikos@telecom.tuc.gr). This author was supported in part by the U.S. Army Research Office (ARO), through its European Research Office (ERO), under ERO contract N62558-03-C-0012 and the European Union under U-BROAD STREP grant 506790.

§Department of Mathematics, University of Washington, Seattle, WA 98195 (tseng@math.washington.edu). This author was supported by National Science Foundation grants DMS-0511283 and DMS-0610037.

¶Department of Systems Engineering and Engineering Management, Chinese University of Hong Kong, Shatin, Hong Kong (zhang@se.cuhk.edu.hk). This author was supported by Hong Kong RGC earmarked grant CUHK418505.

in a region defined by the intersection of the exteriors of m cocentered ellipsoids. If the vectors $h_1, \ldots, h_M$ are linearly independent, then M equals the sum of the ranks of the matrices defining these m ellipsoids. Notice that the problem (1) is easily solved for the case of $n = 1$, so we assume $n \geq 2$.

We assume that $\sum_{\ell \in \mathcal{I}_i} \|h_\ell\| \neq 0$ for all i, which is clearly a necessary condition for (1) to be feasible. This is also a sufficient condition (since $\bigcup_{i=1}^m \{z \mid \sum_{\ell \in \mathcal{I}_i} |h_\ell^H z|^2 = 0\}$ is a finite union of proper subspaces of $\mathbb{F}^n$, so its complement is nonempty and any point in its complement can be scaled to be feasible for (1)). Thus, problem (1) always has an optimal solution (not unique) since its objective function is coercive and continuous, and its feasible set is nonempty and closed. Notice, however, that the feasible set of (1) is typically nonconvex and disconnected, with an exponential number of connected components exhibiting little symmetry. This is in contrast to the quadratic problems with convex feasible set but nonconvex objective function considered in [13, 14, 22]. Furthermore, unlike the class of quadratic problems studied in [1, 7, 8, 15, 16, 21, 23, 24, 25, 26], the constraint functions in (1) do not depend on $z_1^2, \ldots, z_n^2$ only.

Our interest in the nonconvex quadratic program (QP) (1) is motivated by the transmit beamforming problem for multicasting applications [20] and by the wireless sensor network localization problem [6]. In the transmit beamforming problem, a transmitter utilizes an array of n transmitting antennas to broadcast information within its service area to m radio receivers, with receiver $i \in \{1, \ldots, m\}$ equipped with $|\mathcal{I}_i|$ receiving antennas. Let h_ℓ, $\ell \in \mathcal{I}_i$, denote the $n \times 1$ complex *steering vector* modeling propagation loss and phase shift from the transmitting antennas to the ℓth receiving antenna of receiver i. Assuming that each receiver performs spatially matched filtering/maximum ratio combining, which is the optimal combining strategy under standard mild assumptions, then the constraint

$$\sum_{\ell \in \mathcal{I}_i} |h_\ell^H z|^2 \geq 1$$

models the requirement that the total received signal power at receiver i must be above a given threshold (normalized to 1). This constraint is also equivalent to a signal-to-noise ratio (SNR) condition commonly used in data communication. Thus, to minimize the total transmit power subject to individual SNR requirements (one at each receiver), we are led to the QP (1). In the special case where each radio receiver is equipped with a single receiving antenna, the problem reduces to [20]

$$\begin{array}{ll} \min & \|z\|^2 \\ \text{s.t.} & |h_\ell^H z|^2 \geq 1, \quad \ell = 1, \ldots, m, \\ & z \in \mathbb{F}^n. \end{array} \tag{2}$$

This problem is a special case of (1), whereby each ellipsoid lies in $\mathbb{F}^n$ and the corresponding matrix has rank 1.

In this paper, we first show that the nonconvex QP (2) is NP-hard in either the real or the complex case, which further implies the NP-hardness of the general problem (1). Then, we consider a semidefinite programming (SDP) *relaxation* of (1) and a convex QP *restriction* of (2) and study their worst-case performance. In particular, let $v_{\rm sdp}$, $v_{\rm cqp}$, and $v_{\rm qp}$ denote the optimal values of the SDP relaxation, the convex QP restriction, and the original QP (1), respectively. We establish a performance ratio of $v_{\rm qp}/v_{\rm sdp} = O(m^2)$ for the SDP relaxation in the real case, and we give an example showing that this bound is tight up to a constant factor. Similarly, we establish a

performance ratio of $\upsilon_{\rm qp}/\upsilon_{\rm sdp} = O(m)$ in the complex case, and we give an example showing the tightness of this bound. We further show that in the case when the phase spread of the entries of $h_1, \ldots, h_M$ is bounded away from $\pi/2$, the performance ratios $\upsilon_{\rm qp}/\upsilon_{\rm sdp}$ and $\upsilon_{\rm cqp}/\upsilon_{\rm qp}$ for the SDP relaxation and the convex QP restriction, respectively, are independent of m and n.

In recent years, there have been extensive studies of the performance of SDP relaxations for nonconvex QP. However, to our knowledge, this is the first performance analysis of SDP relaxation for QP with concave quadratic constraints. Our proof techniques also extend to a maximization version of the QP (1) with convex homogeneous quadratic constraints. In particular, we give a simple proof of a result analogous to one of Nemirovski et al. [14] (also see [13, Theorem 4.7]) for the real case, namely, the SDP relaxation for this nonconvex QP has a performance ratio of $O(1/\ln(m))$.

2. NP-hardness. In this section, we show that the nonconvex QP (1) is NP-hard in general. First, we notice that, by a linear transformation if necessary, the problem

$$\begin{array}{lll} \text{minimize} & z^H Q z & \\ \text{s.t.} & |z_\ell| \geq 1, & \ell = 1, \ldots, n, \\ & z \in \mathbb{F}^n, & \end{array} \tag{3}$$

is a special case of (1), where $Q \in \mathbb{F}^{n\times n}$ is a Hermitian positive definite matrix (i.e., $Q \succ 0$), and z_ℓ denotes the ℓth component of z. Hence, it suffices to establish the NP-hardness of (3). To this end, we consider a reduction from the NP-complete partition problem: Given positive integers $a_1,\ a_2,\ \ldots,\ a_N$, decide whether there exists a subset $\mathcal{I}$ of $\{1, \ldots, N\}$ satisfying

$$\sum_{\ell\in\mathcal{I}} a_\ell = \frac{1}{2}\sum_{\ell=1}^{N} a_\ell. \tag{4}$$

Our reductions differ for the real and complex cases. As will be seen, the NP-hardness proof in the complex case[1] is more intricate than in the real case.

2.1. The real case. We consider the real case of $\mathbb{F} = \mathbb{R}$. Let $n := N$ and

$$\begin{aligned} a &:= (a_1, \ldots, a_N)^T, \\ Q &:= aa^T + I_n \ \succ\ 0, \end{aligned}$$

where I_n denotes the $n \times n$ identity matrix.

We show that a subset $\mathcal{I}$ satisfying (4) exists if and only if the optimization problem (3) has a minimum value of n. Since

$$z^T Q z = |a^T z|^2 + \sum_{\ell=1}^{n} |z_\ell|^2 \geq n \quad \text{ whenever } \ |z_\ell| \geq 1 \ \forall\ \ell,\ z \in \mathbb{R}^n,$$

we see that (3) has a minimum value of n if and only if there exists a $z \in \mathbb{R}^n$ satisfying

$$a^T z = 0, \quad |z_\ell| = 1 \ \ \forall\ \ell.$$

The above condition is equivalent to the existence of a subset $\mathcal{I}$ satisfying (4), with the correspondence $\mathcal{I} = \{\ell \mid z_\ell = 1\}$. This completes the proof.

[1] This NP-hardness proof was first presented in an appendix of [20] and is included here for completeness; also see [26, Proposition 3.5] for a related proof.

2.2. The complex case. We consider the complex case of $\mathbb{F} = \mathbb{C}$. Let $n := 2N + 1$ and

$$\begin{aligned} a &:= (a_1, \ldots, a_N)^T, \\ A &:= \begin{pmatrix} I_N & I_N & -e_N \\ a^T & 0_N^T & -\frac{1}{2}a^T e_N \end{pmatrix}, \\ Q &:= A^T A + I_n \succ 0, \end{aligned}$$

where e_N denotes the N-dimensional vector of ones, 0_N denotes the N-dimensional vector of zeros, and I_n and I_N are identity matrices of sizes $n \times n$ and $N \times N$, respectively.

We show that a subset $\mathcal{I}$ satisfying (4) exists if and only if the optimization problem (3) has a minimum value of n. Since

$$z^H Q z = \|Az\|^2 + \sum_{\ell=1}^{n} |z_\ell|^2 \geq n \quad \text{whenever } |z_\ell| \geq 1 \ \forall\, \ell,\ z \in \mathbb{C}^n,$$

we see that (3) has a minimum value of n if and only if there exists a $z \in \mathbb{C}^n$ satisfying

$$Az = 0, \quad |z_\ell| = 1 \ \ \forall\, \ell.$$

Expanding $Az = 0$ gives the following set of linear equations:

$$0 = z_\ell + z_{N+\ell} - z_n, \quad \ell = 1, \ldots, N, \tag{5}$$

$$0 = \sum_{\ell=1}^{N} a_\ell z_\ell - \frac{1}{2}\left(\sum_{\ell=1}^{N} a_\ell\right) z_n. \tag{6}$$

For $\ell = 1, \ldots, 2N$, since $|z_\ell| = |z_n| = 1$ so that $z_\ell / z_n = e^{\mathbf{i}\theta_\ell}$ for some $\theta_\ell \in [0, 2\pi)$, we can rewrite (5) as

$$\begin{array}{rcll} \cos\theta_\ell + \cos\theta_{N+\ell} & = & 1, & \\ \sin\theta_\ell + \sin\theta_{N+\ell} & = & 0, & \ell = 1, \ldots, N. \end{array}$$

These equations imply that $\theta_\ell \in \{-\pi/3, \pi/3\}$ for all $\ell \neq n$. In fact, these equations further imply that $\cos\theta_\ell = \cos\theta_{N+\ell} = 1/2$ for $\ell = 1, \ldots, N$, so that

$$\mathrm{Re}\left(\sum_{\ell=1}^{N} a_\ell \frac{z_\ell}{z_n} - \frac{1}{2}\left(\sum_{\ell=1}^{N} a_\ell\right)\right) = 0.$$

Therefore, (6) is satisfied if and only if

$$\mathrm{Im}\left(\sum_{\ell=1}^{N} a_\ell \frac{z_\ell}{z_n} - \frac{1}{2}\left(\sum_{\ell=1}^{N} a_\ell\right)\right) = \mathrm{Im}\left(\sum_{\ell=1}^{N} a_\ell \frac{z_\ell}{z_n}\right) = 0,$$

which is further equivalent to the existence of a subset $\mathcal{I}$ satisfying (4), with the correspondence $\mathcal{I} = \{\ell \mid \theta_\ell = \pi/3\}$. This completes the proof.

3. Performance analysis of SDP relaxation. In this section, we study the performance of an SDP relaxation of (2). Let

$$H_i := \sum_{\ell \in \mathcal{I}_i} h_\ell h_\ell^H, \quad i = 1, \ldots, m.$$

The well-known SDP relaxation of (1) [11, 19] is

$$\begin{array}{rcll} \upsilon_{\text{sdp}} & := & \min & \text{Tr}(Z) \\ & & \text{s.t.} & \text{Tr}(H_i Z) \geq 1, \quad i = 1, \ldots, m, \\ & & & Z \succeq 0, \quad Z \in \mathbb{F}^{n \times n} \text{ is Hermitian.} \end{array} \tag{7}$$

An optimal solution of the SDP relaxation (7) can be computed efficiently using, say, interior-point methods; see [18] and references therein.

Clearly $\upsilon_{\text{sdp}} \leq \upsilon_{\text{qp}}$. We are interested in upper bounds for the relaxation performance of the form

$$\upsilon_{\text{qp}} \leq C \upsilon_{\text{sdp}},$$

where $C \geq 1$. Since we assume $H_i \neq 0$ for all i, it is easily checked that (7) has an optimal solution, which we denote by Z^*.

3.1. General steering vectors: The real case. We consider the real case of $\mathbb{F} = \mathbb{R}$. Upon obtaining an optimal solution Z^* of (7), we construct a feasible solution of (1) using the following randomization procedure:

1. Generate a random vector $\xi \in \mathbb{R}^n$ from the real-valued normal distribution $N(0, Z^*)$.
2. Let $z^*(\xi) = \xi / \min_{1 \leq i \leq m} \sqrt{\xi^T H_i \xi}$.

We will use $z^*(\xi)$ to analyze the performance of the SDP relaxation. Similar procedures have been used for related problems [1, 3, 4, 5, 14]. First, we need to develop two lemmas. The first lemma estimates the left-tail of the distribution of a convex quadratic form of a Gaussian random vector.

LEMMA 1. *Let $H \in \mathbb{R}^{n \times n}$, $Z \in \mathbb{R}^{n \times n}$ be two symmetric positive semidefinite matrices (i.e., $H \succeq 0$, $Z \succeq 0$). Suppose $\xi \in \mathbb{R}^n$ is a random vector generated from the real-valued normal distribution $N(0, Z)$. Then, for any $\gamma > 0$,*

$$\text{Prob}\left(\xi^T H \xi < \gamma E(\xi^T H \xi)\right) \leq \max\left\{\sqrt{\gamma}, \frac{2(\bar{r} - 1)\gamma}{\pi - 2}\right\}, \tag{8}$$

where $\bar{r} := \min\{\text{rank}(H), \text{rank}(Z)\}$.

Proof. Since the covariance matrix $Z \succeq 0$ has rank $r := \text{rank}(Z)$, we can write $Z = UU^T$, for some $U \in \mathbb{R}^{n \times r}$ satisfying $U^T Z U = I_r$. Let $\bar{\xi} := Q^T U^T \xi \in \mathbb{R}^r$, where $Q \in \mathbb{R}^{r \times r}$ is an orthogonal matrix corresponding to the eigen-decomposition of the matrix

$$U^T H U = Q \Lambda Q^T$$

for some diagonal matrix $\Lambda = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_r\}$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r \geq 0$. Since $U^T H U$ has rank at most $\bar{r}$, we have $\lambda_i = 0$ for all $i > \bar{r}$. It is readily checked

that $\bar{\xi}$ has the normal distribution $N(0, I_r)$. Moreover, ξ is statistically identical to $UQ\bar{\xi}$, so that $\xi^T H\xi$ is statistically identical to

$$\bar{\xi}^T Q^T U^T HUQ\bar{\xi} = \bar{\xi}^T \Lambda \bar{\xi} = \sum_{i=1}^{\bar{r}} \lambda_i |\bar{\xi}_i|^2.$$

Then we have

$$\begin{aligned}\text{Prob}\left(\xi^T H\xi < \gamma E(\xi^T H\xi)\right) &= \text{Prob}\left(\sum_{i=1}^{\bar{r}} \lambda_i |\bar{\xi}_i|^2 < \gamma E\left(\sum_{i=1}^{\bar{r}} \lambda_i |\bar{\xi}_i|^2\right)\right) \\ &= \text{Prob}\left(\sum_{i=1}^{\bar{r}} \lambda_i |\bar{\xi}_i|^2 < \gamma \sum_{i=1}^{\bar{r}} \lambda_i\right).\end{aligned}$$

If $\lambda_1 = 0$, then this probability is zero, which proves (8). Thus, we will assume that $\lambda_1 > 0$. Let $\bar{\lambda}_i := \lambda_i/(\lambda_1 + \cdots + \lambda_{\bar{r}})$ for $i = 1, \ldots, \bar{r}$. Clearly, we have

$$\bar{\lambda}_1 + \cdots + \bar{\lambda}_{\bar{r}} = 1, \quad \bar{\lambda}_1 \geq \bar{\lambda}_2 \geq \cdots \geq \bar{\lambda}_{\bar{r}} \geq 0.$$

We consider two cases. First, suppose $\bar{\lambda}_1 \geq \alpha$, where $0 < \alpha < 1$. Then, we can bound the above probability as follows:

$$\begin{aligned}\text{Prob}\left(\xi^T H\xi < \gamma E(\xi^T H\xi)\right) &= \text{Prob}\left(\sum_{i=1}^{\bar{r}} \bar{\lambda}_i |\bar{\xi}_i|^2 < \gamma\right) \\ &\leq \text{Prob}\left(\bar{\lambda}_1 |\bar{\xi}_1|^2 < \gamma\right) \\ &\leq \text{Prob}\left(|\bar{\xi}_1|^2 < \gamma/\alpha\right) \\ &\leq \sqrt{\frac{2\gamma}{\pi\alpha}},\end{aligned} \tag{9}$$

where the last step is due to the fact that $\bar{\xi}_1$ is a real-valued zero mean Gaussian random variable with unit variance.

In the second case, we have $\bar{\lambda}_1 < \alpha$, so that

$$\bar{\lambda}_2 + \cdots + \bar{\lambda}_{\bar{r}} = 1 - \bar{\lambda}_1 > 1 - \alpha.$$

This further implies $(\bar{r} - 1)\bar{\lambda}_2 \geq \bar{\lambda}_2 + \cdots + \bar{\lambda}_{\bar{r}} > 1 - \alpha$. Hence

$$\bar{\lambda}_1 \geq \bar{\lambda}_2 > \frac{1-\alpha}{\bar{r}-1}.$$

Using this bound, we obtain the following probability estimate:

$$\begin{aligned}\text{Prob}\left(\xi^T H\xi < \gamma E(\xi^T H\xi)\right) &= \text{Prob}\left(\sum_{i=1}^{\bar{r}} \bar{\lambda}_i |\bar{\xi}_i|^2 < \gamma\right) \\ &\leq \text{Prob}\left(\bar{\lambda}_1 |\bar{\xi}_1|^2 < \gamma,\ \bar{\lambda}_2 |\bar{\xi}_2|^2 < \gamma\right) \\ &= \text{Prob}\left(\bar{\lambda}_1 |\bar{\xi}_1|^2 < \gamma\right) \cdot \text{Prob}\left(\bar{\lambda}_2 |\bar{\xi}_2|^2 < \gamma\right) \\ &\leq \sqrt{\frac{2\gamma}{\pi\bar{\lambda}_1}} \cdot \sqrt{\frac{2\gamma}{\pi\bar{\lambda}_2}} \\ &\leq \frac{2(\bar{r}-1)\gamma}{\pi(1-\alpha)}.\end{aligned} \tag{10}$$

Combining the estimates for the above two cases and setting $\alpha = 2/\pi$, we immediately obtain the desired bound (8). ∎

LEMMA 2. *Let $\mathbb{F} = \mathbb{R}$. Let $Z^* \succeq 0$ be a feasible solution of* (7) *and let $z^*(\xi)$ be generated by the randomization procedure described earlier. Then, with probability* 1, *$z^*(\xi)$ is well defined and feasible for* (1). *Moreover, for every $\gamma > 0$ and $\mu > 0$,*

(11)
$$\text{Prob}\left(\min_{1\le i\le m} \xi^T H_i \xi \ge \gamma,\ \|\xi\|^2 \le \mu \text{Tr}(Z^*)\right) \ge 1 - m \cdot \max\left\{\sqrt{\gamma}, \frac{2(r-1)\gamma}{\pi - 2}\right\} - \frac{1}{\mu},$$

where $r := \text{rank}(Z^*)$.

Proof. Since $Z^* \succeq 0$ is feasible for (7), it follows that $\text{Tr}(H_i Z^*) \ge 1$ for all $i = 1, \ldots, m$. Since $E(\xi^T H_i \xi) = \text{Tr}(H_i Z^*) \ge 1$ and the density of $\xi^T H_i \xi$ is absolutely continuous, the probability of $\xi^T H_i \xi = 0$ is zero, implying that $z^*(\xi)$ is well defined with probability 1. The feasibility of $z^*(\xi)$ is easily verified.

To prove (11), we first note that $E(\xi\xi^T) = Z^*$. Thus, for any $\gamma > 0$ and $\mu > 0$,

$$\begin{aligned}
&\text{Prob}\left(\min_{1\le i\le m} \xi^T H_i \xi \ge \gamma,\ \|\xi\|^2 \le \mu \text{Tr}(Z^*)\right) \\
&\quad = \text{Prob}\left(\xi^T H_i \xi \ge \gamma\ \forall\, i = 1, \ldots, m \ \text{ and } \ \|\xi\|^2 \le \mu \text{Tr}(Z^*)\right) \\
&\quad \ge \text{Prob}\left(\xi^T H_i \xi \ge \gamma \text{Tr}(H_i Z^*)\ \forall\, i = 1, \ldots, m \ \text{ and } \ \|\xi\|^2 \le \mu \text{Tr}(Z^*)\right) \\
&\quad = \text{Prob}\left(\xi^T H_i \xi \ge \gamma E(\xi^T H_i \xi)\ \forall\, i = 1, \ldots, m \ \text{ and } \ \|\xi\|^2 \le \mu E(\|\xi\|^2)\right) \\
&\quad = 1 - \text{Prob}\left(\xi^T H_i \xi < \gamma E(\xi^T H_i \xi) \text{ for some } i \text{ or } \|\xi\|^2 > \mu E(\|\xi\|^2)\right) \\
&\quad \ge 1 - \sum_{i=1}^m \text{Prob}\left(\xi^T H_i \xi < \gamma E(\xi^T H_i \xi)\right) - \text{Prob}\left(\|\xi\|^2 > \mu E(\|\xi\|^2)\right) \\
&\quad > 1 - m \cdot \max\left\{\sqrt{\gamma}, \frac{2(r-1)\gamma}{\pi - 2}\right\} - \frac{1}{\mu},
\end{aligned}$$

where the last step uses Lemma 1 as well as Markov's inequality:

$$\text{Prob}\left(\|\xi\|^2 > \mu E(\|\xi\|^2)\right) \le \frac{1}{\mu}.$$

This completes the proof. ∎

We now use Lemma 2 to bound the performance of the SDP relaxation.

THEOREM 1. *Let $\mathbb{F} = \mathbb{R}$. For the QP* (1) *and its SDP relaxation* (7), *we have $\upsilon_{\text{qp}} = \upsilon_{\text{sdp}}$ if $m \le 2$, and otherwise*

$$\upsilon_{\text{qp}} \le \frac{27m^2}{\pi} \upsilon_{\text{sdp}}.$$

Proof. By applying a suitable rank reduction procedure if necessary, we can assume that the rank r of the optimal SDP solution Z^* satisfies $r(r+1)/2 \le m$; see, e.g., [17]. Thus $r < \sqrt{2m}$. If $m \le 2$, then $r = 1$, implying that $Z^* = z^*(z^*)^T$ for some $z^* \in \mathbb{R}^n$, and it is readily seen that z^* is an optimal solution of (1), so that $\upsilon_{\text{qp}} = \upsilon_{\text{sdp}}$. Otherwise, we apply the randomization procedure to Z^*. We also choose

$$\mu = 3, \quad \gamma = \frac{\pi}{4m^2}\left(1 - \frac{1}{\mu}\right)^2 = \frac{\pi}{9m^2}.$$

Then, it is easily verified using $r < \sqrt{2m}$ that

$$\sqrt{\gamma} \geq \frac{2(r-1)\gamma}{\pi - 2} \quad \forall\ m = 1, 2, \ldots .$$

Plugging these choices of γ and μ into (11), we see that there is a positive probability (independent of problem size) of at least

$$1 - m\sqrt{\gamma} - \frac{1}{\mu} = 1 - \frac{\sqrt{\pi}}{3} - \frac{1}{3} = 0.0758\ldots$$

that ξ generated by the randomization procedure satisfies

$$\min_{1 \leq i \leq m} \xi^T H_i \xi \geq \frac{\pi}{9m^2} \quad \text{and} \quad \|\xi\|^2 \leq 3\,\mathrm{Tr}(Z^*).$$

Let ξ be any vector satisfying these two conditions.[2] Then, $z^*(\xi)$ is feasible for (1), so that

$$\upsilon_{\mathrm{qp}} \leq \|z^*(\xi)\|^2 = \frac{\|\xi\|^2}{\min_i \xi^T H_i \xi} \leq \frac{3\,\mathrm{Tr}(Z^*)}{(\pi/9m^2)} = \frac{27m^2}{\pi}\,\upsilon_{\mathrm{sdp}},$$

where the last equality uses $\mathrm{Tr}(Z^*) = \upsilon_{\mathrm{sdp}}$. □

In the above proof, other choices of μ can also be used, but the resulting bound seems not as sharp. Theorem 1 suggests that the worst-case performance of the SDP relaxation deteriorates quadratically with the number of quadratic constraints. Below we give an example demonstrating that this bound is in fact tight up to a constant factor.

Example 1. For any $m \geq 2$ and $n \geq 2$, consider a special instance of (2), corresponding to (1) with $|\mathcal{I}_i| = 1$ (i.e., each H_i has rank 1), whereby

$$h_\ell = \left(\cos\left(\frac{\ell\pi}{m}\right), \sin\left(\frac{\ell\pi}{m}\right), 0, \ldots, 0\right)^T, \qquad \ell = 1, \ldots, m.$$

Let $z^* = (z_1^*, \ldots, z_n^*)^T \in \mathbb{R}^n$ be an optimal solution of (2) corresponding to the above choice of steering vectors h_ℓ. We can write

$$(z_1^*, z_2^*) = \rho(\cos\theta, \sin\theta) \quad \text{for some } \theta \in [0, 2\pi).$$

Since $\{\ell\pi/m,\ \ell = 1, \ldots, m\}$ is uniformly spaced on $[0, \pi)$, there must exist an integer ℓ such that

$$\text{either} \quad \left|\theta - \frac{\ell\pi}{m} - \frac{\pi}{2}\right| \leq \frac{\pi}{2m} \quad \text{or} \quad \left|\theta - \frac{\ell\pi}{m} + \frac{\pi}{2}\right| \leq \frac{\pi}{2m}.$$

For simplicity, we assume the first case. (The second case can be treated similarly.) Since the last $(n-2)$ entries of h_ℓ are zero, it is readily checked that

$$|h_\ell^T z^*| = \rho\left|\cos\left(\theta - \frac{\ell\pi}{m}\right)\right| = \rho\left|\sin\left(\theta - \frac{\ell\pi}{m} - \frac{\pi}{2}\right)\right| \leq \rho\left|\sin\left(\frac{\pi}{2m}\right)\right| \leq \frac{\rho\pi}{2m}.$$

[2] The probability that no such ξ is generated after N independent trials is at most $(1-0.0758..)^N$, which for $N = 100$ equals 0.000375.. Thus, such ξ requires relatively few trials to generate.

Since z^* satisfies the constraint $|h_\ell^T z^*| \geq 1$, it follows that

$$\|z^*\| \geq \rho \geq \frac{2m|h_\ell^T z^*|}{\pi} \geq \frac{2m}{\pi},$$

implying

$$\upsilon_{\rm qp} = \|z^*\|^2 \geq \frac{4m^2}{\pi^2}.$$

On the other hand, the positive semidefinite matrix

$$Z^* = \mathrm{diag}\{1, 1, 0, \ldots, 0\}$$

is feasible for the SDP relaxation (7), and it has an objective value of $\mathrm{Tr}(Z^*) = 2$. Thus, for this instance, we have

$$\upsilon_{\rm qp} \geq \frac{2m^2}{\pi^2}\, \upsilon_{\rm sdp}.$$

The preceding example and Theorem 1 show that the SDP relaxation (7) can be weak if the number of quadratic constraints is large, especially when the steering vectors h_ℓ are in a certain sense "uniformly distributed" in space.

3.2. General steering vectors: The complex case. We consider the complex case of $\mathbb{F} = \mathbb{C}$. We will show that the performance ratio of the SDP relaxation (7) improves to $O(m)$ in the complex case (as opposed to $O(m^2)$ in the real case). Similar to the real case, upon obtaining an optimal solution Z^* of (7), we construct a feasible solution of (1) using the following randomization procedure:

1. Generate a random vector $\xi \in \mathbb{C}^n$ from the *complex-valued* normal distribution $N_c(0, Z^*)$ [2, 26].
2. Let $z^*(\xi) = \xi / \min_{1\leq i\leq m} \sqrt{\xi^H H_i \xi}$.

Most of the ensuing performance analysis is similar to that of the real case. In particular, we will also need the following two lemmas analogous to Lemmas 1 and 2.

LEMMA 3. *Let $H \in \mathbb{C}^{n\times n}$, $Z \in \mathbb{C}^{n\times n}$ be two Hermitian positive semidefinite matrices (i.e., $H \succeq 0$, $Z \succeq 0$). Suppose $\xi \in \mathbb{C}^n$ is a random vector generated from the complex-valued normal distribution $N_c(0, Z)$. Then, for any $\gamma > 0$,*

$$\mathrm{Prob}\left(\xi^H H\xi < \gamma E(\xi^H H\xi)\right) \leq \max\left\{\frac{4}{3}\gamma, 16(\bar{r}-1)^2\gamma^2\right\}, \tag{12}$$

where $\bar{r} := \min\{\mathrm{rank}(H), \mathrm{rank}(Z)\}$.

Proof. We follow the same notations and proof as for Lemma 1, except for two blanket changes:

$$\begin{array}{rcl} \text{matrix transpose} & \rightarrow & \text{Hermitian transpose}, \\ \text{orthogonal matrix} & \rightarrow & \text{unitary matrix}. \end{array}$$

Also, $\bar{\xi}$ has the complex-valued normal distribution $N_c(0, I_r)$. With these changes, we consider the same two cases: $\bar{\lambda}_1 \geq \alpha$ and $\bar{\lambda}_1 < \alpha$, where $0 < \alpha < 1$. In the first case, we have similar to (9) that

$$\mathrm{Prob}\left(\xi^H H\xi < \gamma E(\xi^H H\xi)\right) \leq \mathrm{Prob}\left(|\bar{\xi}_1|^2 < \gamma/\alpha\right). \tag{13}$$

Recall that the density function of a complex-valued circular normal random variable $u \sim N_c(0, \sigma^2)$, where σ is the standard deviation, is

$$\frac{1}{\pi\sigma^2} e^{-\frac{|u|^2}{\sigma^2}} \quad \forall\ u \in \mathbb{C}.$$

In polar coordinates, the density function can be written as

$$f(\rho, \theta) = \frac{\rho}{\pi\sigma^2} e^{-\frac{\rho^2}{\sigma^2}} \quad \forall\ \rho \in [0, +\infty),\ \theta \in [0, 2\pi).$$

In fact, a complex-valued normal distribution can be viewed as a joint distribution of its modulus and its argument, with the following particular properties: (1) the modulus and argument are independently distributed; (2) the argument is uniformly distributed over $[0, 2\pi)$; (3) the modulus follows a Weibull distribution with density

$$f(\rho) = \begin{cases} \frac{2\rho}{\sigma^2} e^{-\frac{\rho^2}{\sigma^2}} & \text{if } \rho \geq 0; \\ 0 & \text{if } \rho < 0, \end{cases}$$

and distribution function

$$\text{Prob}\{|u| \leq t\} = 1 - e^{-\frac{t^2}{\sigma^2}}. \tag{14}$$

Since $\bar{\xi}_1 \sim N_c(0, 1)$, substituting this into (13) yields

$$\text{Prob}\left(\xi^H H\xi < \gamma E(\xi^H H\xi)\right) \leq \text{Prob}\left(|\bar{\xi}_1|^2 < \gamma/\alpha\right) \leq 1 - e^{-\gamma/\alpha} \leq \gamma/\alpha,$$

where the last inequality uses the convexity of the exponential function.

In the second case of $\bar{\lambda}_1 < \alpha$, we have similar to (10) that

$$\begin{aligned} \text{Prob}\left(\xi^H H\xi < \gamma E(\xi^H H\xi)\right) &\leq \text{Prob}\left(\bar{\lambda}_1|\bar{\xi}_1|^2 < \gamma\right) \cdot \text{Prob}\left(\bar{\lambda}_2|\bar{\xi}_2|^2 < \gamma\right) \\ &= (1 - e^{-\gamma/\bar{\lambda}_1})(1 - e^{-\gamma/\bar{\lambda}_2}) \\ &\leq \frac{\gamma^2}{\bar{\lambda}_1\bar{\lambda}_2} \\ &\leq \frac{(\bar{r} - 1)^2\gamma^2}{(1 - \alpha)^2}, \end{aligned}$$

where the last step uses the fact that $\bar{\lambda}_1 \geq \bar{\lambda}_2 \geq (1 - \alpha)/(\bar{r} - 1)$. Combining the estimates for the above two cases and setting $\alpha = 3/4$, we immediately obtain the desired bound (12). □

LEMMA 4. *Let $\mathbb{F} = \mathbb{C}$. Let $Z^* \succeq 0$ be a feasible solution of (7) and let $z^*(\xi)$ be generated by the randomization procedure described earlier. Then, with probability 1, $z^*(\xi)$ is well defined and feasible for (1). Moreover, for every $\gamma > 0$ and $\mu > 0$,*

$$\text{Prob}\left(\min_{1 \leq i \leq m} \xi^H H_i \xi \geq \gamma,\ \|\xi\|^2 \leq \mu \text{Tr}(Z^*)\right) \geq 1 - m \cdot \max\left\{\frac{4}{3}\gamma, 16(r-1)^2\gamma^2\right\} - \frac{1}{\mu},$$

where $r := \text{rank}(Z^*)$.

Proof. The proof is mostly the same as that for the real case (see Lemma 2). In particular, for any $\gamma > 0$ and $\mu > 0$, we still have

$$\begin{aligned} &\text{Prob}\left(\min_{1 \leq i \leq m} \xi^H H_i \xi \geq \gamma,\ \|\xi\|^2 \leq \mu \text{Tr}(Z^*)\right) \\ &\quad \geq 1 - \sum_{i=1}^{m} \text{Prob}\left(\xi^H H_i \xi < \gamma E(\xi^H H_i \xi)\right) - \text{Prob}\left(\|\xi\|^2 > \mu E(\|\xi\|^2)\right). \end{aligned}$$

Therefore, we can invoke Lemma 3 to obtain

$$\begin{aligned}
&\mathrm{Prob}\left(\min_{1\le i\le m} \xi^H H_i \xi \ge \gamma,\ \|\xi\|^2 \le \mu \mathrm{Tr}(Z^*)\right) \\
&\quad \ge 1 - m\cdot\max\left\{\frac{4}{3}\gamma, 16(r-1)^2\gamma^2\right\} - \mathrm{Prob}\left(\|\xi\|^2 > \mu E(\|\xi\|^2)\right) \\
&\quad \ge 1 - m\cdot\max\left\{\frac{4}{3}\gamma, 16(r-1)^2\gamma^2\right\} - \frac{1}{\mu},
\end{aligned}$$

which completes the proof. □

THEOREM 2. *Let $\mathbb{F} = \mathbb{C}$. For the QP* (1) *and its SDP relaxation* (7)*, we have $v_{\mathrm{sdp}} = v_{\mathrm{qp}}$ if $m \le 3$ and otherwise*

$$v_{\mathrm{qp}} \le 8m \cdot v_{\mathrm{sdp}}.$$

Proof. By applying a suitable rank reduction procedure if necessary, we can assume that the rank r of the optimal SDP solution Z^* satisfies $r = 1$ if $m \le 3$ and $r \le \sqrt{m}$ if $m \ge 4$; see [9, section 5]. Thus, if $m \le 3$, then $Z^* = z^*(z^*)^H$ for some $z^* \in \mathbb{C}^n$ and it is readily seen that z^* is an optimal solution of (1), so that $v_{\mathrm{sdp}} = v_{\mathrm{qp}}$. Otherwise, we apply the randomization procedure to Z^*. By choosing $\mu = 2$ and $\gamma = \frac{1}{4m}$, it is easily verified using $r \le \sqrt{m}$ that

$$\frac{4}{3}\gamma \ge 16(r-1)^2\gamma^2 \quad \forall\ m = 1, 2, \ldots .$$

Therefore, it follows from Lemma 4 that

$$\mathrm{Prob}\left\{\min_{1\le i\le m} \xi^H H_i \xi \ge \gamma,\ \|\xi\|^2 \le \mu \mathrm{Tr}(Z^*)\right\} \ge 1 - m\frac{4}{3}\gamma - \frac{1}{\mu} = \frac{1}{6}.$$

Then, similar to the proof of Theorem 1, we obtain that with probability of at least 1/6, $z^*(\xi)$ is a feasible solution of (1) and $v_{\mathrm{qp}} \le \|z^*(\xi)\|^2 \le 8m \cdot v_{\mathrm{sdp}}$.[3] □

The proof of Theorem 2 shows that by repeating the randomization procedure, the probability of generating a feasible solution with a performance ratio no more than $8m$ approaches 1 exponentially fast (independent of problem size). Alternatively, a derandomization technique from theoretical computer science can perhaps convert the above randomization procedure into a polynomial-time deterministic algorithm [12]; also see [14].

Theorem 2 shows that the worst-case performance of SDP relaxation deteriorates *linearly* with the number of quadratic constraints. This contrasts with the *quadratic* rate of deterioration in the real case (see Theorem 1). Thus, the SDP relaxation can yield better performance in the complex case. This is in the same spirit as the recent results in [26] which showed that the quality of SDP relaxation improves by a constant factor for certain quadratic *maximization* problems when the space is changed from $\mathbb{R}^n$ to $\mathbb{C}^n$. Below we give an example demonstrating that this approximation bound is tight up to a constant factor.

Example 2. For any $m \ge 2$ and $n \ge 2$, let $K = \lceil \sqrt{m} \rceil$ (so $K \ge 2$). Consider a special instance of (2), corresponding to (1) with $|\mathcal{I}_i| = 1$ (i.e., each H_i has rank 1),

[3] The probability that no such ξ is generated after N independent trials is at most $(5/6)^N$, which for $N = 30$ equals 0.00421.. Thus, such ξ requires relatively few trials to generate.

whereby

$$h_\ell = \left(\cos\frac{j\pi}{K}, \sin\frac{j\pi}{K} e^{\mathbf{i}\frac{2k\pi}{K}}, 0, \ldots, 0\right)^T \quad \text{with} \;\; \ell = jK - K + k, \qquad j,k = 1,\ldots,K.$$

Hence there are K^2 complex rank-1 constraints. Let $z^* = (z_1^*, \ldots, z_n^*)^T \in \mathbb{C}^n$ be an optimal solution of (2) corresponding to the above choice of $\lceil\sqrt{m}\rceil^2$ steering vectors h_ℓ. By a phase rotation if necessary, we can without loss of generality assume that z_1^* is real and write

$$(z_1^*, z_2^*) = \rho(\cos\theta, \sin\theta e^{\mathbf{i}\psi}) \quad \text{for some } \theta, \psi \in [0, 2\pi).$$

Since $\{2k\pi/K,\ k = 1,\ldots,K\}$ and $\{j\pi/K,\ j = 1,\ldots,K\}$ are uniformly spaced in $[0, 2\pi)$ and $[0, \pi)$, respectively, there must exist integers j and k such that

$$\left|\psi - \frac{2k\pi}{K}\right| \le \frac{\pi}{K} \quad \text{and} \quad \text{either } \left|\theta - \frac{j\pi}{K} - \frac{\pi}{2}\right| \le \frac{\pi}{2K} \text{ or } \left|\theta - \frac{j\pi}{K} + \frac{\pi}{2}\right| \le \frac{\pi}{2K}.$$

Without loss of generality, we assume

$$\left|\theta - \frac{j\pi}{K} - \frac{\pi}{2}\right| \le \frac{\pi}{2K}.$$

Since the last $(n-2)$ entries of each h_ℓ are zero, it is readily seen that for $\ell = jK-K+k$,

$$\begin{aligned}
\left|\mathrm{Re}(h_\ell^H z^*)\right| &= \rho\left|\cos\theta\cos\frac{j\pi}{K} + \sin\theta\sin\frac{j\pi}{K}\cos\left(\psi - \frac{2k\pi}{K}\right)\right| \\
&= \rho\left|\cos\left(\theta - \frac{j\pi}{K}\right) + \sin\theta\sin\frac{j\pi}{K}\left(\cos\left(\psi - \frac{2k\pi}{K}\right) - 1\right)\right| \\
&= \rho\left|\sin\left(\theta - \frac{j\pi}{K} - \frac{\pi}{2}\right) - 2\sin\theta\sin\frac{j\pi}{K}\sin^2\left(\frac{K\psi - 2k\pi}{2K}\right)\right| \\
&\le \rho\left|\sin\frac{\pi}{2K}\right| + 2\rho\sin^2\frac{\pi}{2K} \\
&\le \frac{\rho\pi}{2K} + \frac{\rho\pi^2}{2K^2}.
\end{aligned}$$

In addition, we have

$$\begin{aligned}
\left|\mathrm{Im}(h_\ell^H z^*)\right| &= \rho\left|\sin\theta\sin\frac{j\pi}{K}\sin\left(\psi - \frac{2k\pi}{K}\right)\right| \\
&\le \rho\left|\sin\left(\psi - \frac{2k\pi}{K}\right)\right| \\
&\le \rho\left|\psi - \frac{2k\pi}{K}\right| \le \frac{\rho\pi}{K}.
\end{aligned}$$

Combining the above two bounds, we obtain

$$\left|h_\ell^H z^*\right| \le \left|\mathrm{Re}(h_\ell^H z^*)\right| + \left|\mathrm{Im}(h_\ell^H z^*)\right| \le \frac{3\rho\pi}{2K} + \frac{\rho\pi^2}{2K^2}.$$

Since z^* satisfies the constraint $|h_\ell^H z^*| \ge 1$, it follows that

$$\|z^*\| \ge \rho \ge \frac{2K^2|h_\ell^H z^*|}{\pi(3K+\pi)} \ge \frac{2K^2}{\pi(3K+\pi)},$$

implying

$$v_{\rm qp} = \|z^*\|^2 \geq \frac{4K^4}{\pi^2(3K+\pi)^2} = \frac{4\lceil\sqrt{m}\rceil^4}{\pi^2(3\lceil\sqrt{m}\rceil+\pi)^2}.$$

On the other hand, the positive semidefinite matrix

$$Z^* = \mathrm{diag}\{1, 1, 0, \ldots, 0\}$$

is feasible for the SDP relaxation (7), and it has an objective value of $\mathrm{Tr}(Z^*) = 2$. Thus, for this instance, we have

$$v_{\rm qp} \geq \frac{2\lceil\sqrt{m}\rceil^4}{\pi^2(3\lceil\sqrt{m}\rceil+\pi)^2}\ v_{\rm sdp} \geq \frac{2m}{\pi^2(3+\pi/2)^2}\ v_{\rm sdp}.$$

The preceding example and Theorem 2 show that the SDP relaxation (7) can be weak if the number of quadratic constraints is large, especially when the steering vectors h_ℓ are in a certain sense uniformly distributed in space. In the next subsection, we will tighten the approximation bound in Theorem 2 by considering special cases where the steering vectors are not too spread out in space.

3.3. Specially configured steering vectors: The complex case. We consider the complex case of $\mathbb{F} = \mathbb{C}$. Let Z^* be any optimal solution of (7). Since Z^* is feasible for (7), $Z^* \neq 0$. Then

$$Z^* = \sum_{k=1}^{r} w_k w_k^H \tag{15}$$

for some nonzero $w_k \in \mathbb{C}^n$, where $r := \mathrm{rank}(Z^*) \geq 1$. By decomposing $w_k = u_k + v_k$, with $u_k \in \mathrm{span}\{h_1, \ldots, h_M\}$ and $v_k \in \mathrm{span}\{h_1, \ldots, h_M\}^\perp$, it is easily checked that $\tilde{Z} := \sum_{k=1}^r u_k u_k^H$ is feasible for (7) and

$$\langle I, Z^*\rangle = \sum_{k=1}^{r} \|u_k + v_k\|^2 = \sum_{k=1}^{r}(\|u_k\|^2 + \|v_k\|^2) = \langle I, \tilde{Z}\rangle + \sum_{k=1}^{r}\|v_k\|^2.$$

This implies $v_k = 0$ for all k, so that

$$w_k \in \mathrm{span}\{h_1, \ldots, h_M\}. \tag{16}$$

Below we show that the SDP relaxation (7) provides a *constant factor approximation* to the QP (1) when the phase spread of the entries of h_ℓ is bounded away from $\pi/2$.

THEOREM 3. *Suppose that*

$$h_\ell = \sum_{i=1}^{p} \beta_{i\ell} g_i \quad \forall\ \ell = 1, \ldots, M \tag{17}$$

for some $p \geq 1$, $\beta_{i\ell} \in \mathbb{C}$, and $g_i \in \mathbb{C}^n$ such that $\|g_i\| = 1$ and $g_i^H g_j = 0$ for all $i \neq j$. Then the following results hold:

(a) *If $\mathrm{Re}(\beta_{i\ell}^H \beta_{j\ell}) > 0$ whenever $\beta_{i\ell}^H \beta_{j\ell} \neq 0$, then $v_{\rm qp} \leq C v_{\rm sdp}$, where*

$$C := \max_{i,j,\ell\ |\ \beta_{i\ell}^H\beta_{j\ell}\neq 0} \left(1 + \frac{|\mathrm{Im}(\beta_{i\ell}^H\beta_{j\ell})|^2}{|\mathrm{Re}(\beta_{i\ell}^H\beta_{j\ell})|^2}\right)^{1/2}. \tag{18}$$

(b) *If* $\beta_{i\ell} = |\beta_{i\ell}|e^{\mathbf{i}\phi_{i\ell}}$, *where*

$$\phi_{i\ell} \in [\bar{\phi}_\ell - \phi, \bar{\phi}_\ell + \phi] \quad \forall\, i, \ell \ \text{ for some } 0 \le \phi < \frac{\pi}{4} \text{ and some } \bar{\phi}_\ell \in \mathbb{R}, \tag{19}$$

then $\mathrm{Re}(\beta_{i\ell}^H \beta_{j\ell}) > 0$ *whenever* $\beta_{i\ell}^H \beta_{j\ell} \neq 0$, *and* C *given by* (18) *satisfies*

$$C \le \frac{1}{\cos(2\phi)}. \tag{20}$$

Proof. (a) By (16), we have $w_k = \sum_{i=1}^p \alpha_{ki} g_i$ for some $\alpha_{ki} \in \mathbb{C}$. This together with (15) yields

$$\begin{aligned}\langle I, Z^* \rangle = \sum_{k=1}^r \|w_k\|^2 \ &= \ \sum_{k=1}^r \left\| \sum_{i=1}^p \alpha_{ki} g_i \right\|^2 \\ &= \sum_{k=1}^r \sum_{i=1}^p |\alpha_{ki}|^2 \ = \ \sum_{i=1}^p \lambda_i^2,\end{aligned}$$

where the third equality uses the orthonormal properties of $g_1, \ldots, g_p$, and the last equality uses $\lambda_i := \left(\sum_{k=1}^r |\alpha_{ki}|^2\right)^{1/2} = \|(\alpha_{ki})_{k=1}^r\|$.

Let $z^* := \sum_{i=1}^p \lambda_i g_i$. Then, the orthonormal properties of $g_1, \ldots, g_p$ yield

$$\|z^*\|^2 = \left\| \sum_{i=1}^p \lambda_i g_i \right\|^2 = \sum_{i=1}^p \lambda_i^2 = \langle I, Z^* \rangle = \upsilon_{\rm sdp}. \tag{21}$$

Moreover, for each $\ell \in \{1, \ldots, M\}$, we obtain from (15) that

$$\begin{aligned}\langle h_\ell h_\ell^H, Z^* \rangle &= \sum_{k=1}^r \langle h_\ell h_\ell^H, w_k w_k^H \rangle = \sum_{k=1}^r |h_\ell^H w_k|^2 \\ &= \sum_{k=1}^r \left| \sum_{i=1}^p \alpha_{ki} h_\ell^H g_i \right|^2 = \sum_{k=1}^r \left| \sum_{i=1}^p \alpha_{ki} \beta_{i\ell} \right|^2 \\ &= \mathrm{Re}\left(\sum_{k=1}^r \sum_{i=1}^p \sum_{j=1}^p \alpha_{ki}^H \alpha_{kj} \beta_{i\ell}^H \beta_{j\ell} \right) = \mathrm{Re}\left(\sum_{i=1}^p \sum_{j=1}^p \beta_{i\ell}^H \beta_{j\ell} \sum_{k=1}^r \alpha_{ki}^H \alpha_{kj} \right) \\ &= \sum_{i=1}^p \sum_{j=1}^p \mathrm{Re}\left(\beta_{i\ell}^H \beta_{j\ell} \sum_{k=1}^r \alpha_{ki}^H \alpha_{kj} \right) \\ &\le \sum_{i=1}^p \sum_{j=1}^p \left| \beta_{i\ell}^H \beta_{j\ell} \right| \left| \sum_{k=1}^r \alpha_{ki}^H \alpha_{kj} \right| \le \sum_{i=1}^p \sum_{j=1}^p \left| \beta_{i\ell}^H \beta_{j\ell} \right| \|(\alpha_{ki})_{k=1}^r\| \|(\alpha_{kj})_{k=1}^r\| \\ &= \sum_{i=1}^p \sum_{j=1}^p \left| \beta_{i\ell}^H \beta_{j\ell} \right| \lambda_i \lambda_j,\end{aligned}$$

where the fourth equality uses (17) and the orthonormal properties of $g_1, \ldots, g_p$; the last inequality is due to the Cauchy–Schwarz inequality. Then, it follows that

$$\begin{aligned}\langle h_\ell h_\ell^H, Z^* \rangle &\le \sum_{i=1}^p \sum_{j=1}^p \left(|\mathrm{Re}(\beta_{i\ell}^H \beta_{j\ell})|^2 + |\mathrm{Im}(\beta_{i\ell}^H \beta_{j\ell})|^2 \right)^{1/2} \lambda_i \lambda_j \\ &= \sum_{i=1}^p \sum_{j=1}^p \left| \mathrm{Re}(\beta_{i\ell}^H \beta_{j\ell}) \right| \left(1 + \frac{|\mathrm{Im}(\beta_{i\ell}^H \beta_{j\ell})|^2}{|\mathrm{Re}(\beta_{i\ell}^H \beta_{j\ell})|^2} \right)^{1/2} \lambda_i \lambda_j \\ &\le \sum_{i=1}^p \sum_{j=1}^p \left| \mathrm{Re}(\beta_{i\ell}^H \beta_{j\ell}) \right| C \lambda_i \lambda_j \\ &= \sum_{i=1}^p \sum_{j=1}^p \mathrm{Re}(\beta_{i\ell}^H \beta_{j\ell}) C \lambda_i \lambda_j,\end{aligned}$$

where the summation in the second step is taken over i, j with $\beta_{i\ell}^H \beta_{j\ell} \neq 0$, the third step is due to (18), and the last step is due to the assumption that $\mathrm{Re}(\beta_{i\ell}^H \beta_{j\ell}) > 0$ whenever $\beta_{i\ell}^H \beta_{j\ell} \neq 0$. Also, we have from (17) and the orthonormal properties of $g_1, \ldots, g_p$ that

$$|h_\ell^H z^*|^2 = \left\| \sum_{i=1}^p \lambda_i h_\ell^H g_i \right\|^2 = \left\| \sum_{i=1}^p \lambda_i \beta_{i\ell} \right\|^2 = \sum_{i=1}^p \sum_{j=1}^p \lambda_i \lambda_j \mathrm{Re}(\beta_{i\ell}^H \beta_{j\ell}).$$

Comparing the above two displayed equations, we see that

$$\langle h_\ell h_\ell^H, Z^* \rangle \le C |h_\ell^H z^*|^2, \quad \ell = 1, \ldots, M.$$

Since Z^* is feasible for (7), this shows that $\sqrt{C} z^*$ is feasible for (1), which further implies

$$v_{\mathrm{qp}} \le \left\| \sqrt{C} z^* \right\|^2 = C \|z^*\|^2 = C v_{\mathrm{sdp}}.$$

This proves the desired result.

(b) The condition (19) implies that $|\phi_{i\ell} - \phi_{j\ell}| \le 2\phi < \pi/2$. In other words, the phase angle spread of the entries of each $\beta_\ell = (\beta_{1\ell}, \beta_{2\ell}, \ldots, \beta_{n\ell})^T$ is no more than 2ϕ. This further implies that

$$\cos(\phi_{i\ell} - \phi_{j\ell}) \ge \cos(2\phi) \quad \forall\, i, j, \ell. \tag{22}$$

We have

$$\begin{aligned}\beta_{i\ell}^H \beta_{j\ell} &= |\beta_{i\ell}| e^{-\mathbf{i}\phi_{i\ell}} |\beta_{j\ell}| e^{\mathbf{i}\phi_{j\ell}} \\ &= |\beta_{i\ell}||\beta_{j\ell}| e^{\mathbf{i}(\phi_{j\ell} - \phi_{i\ell})} \\ &= |\beta_{i\ell}||\beta_{j\ell}| (\cos(\phi_{j\ell} - \phi_{i\ell}) + \mathbf{i} \sin(\phi_{j\ell} - \phi_{i\ell})).\end{aligned}$$

Since $|\phi_{i\ell} - \phi_{j\ell}| < \pi/2$ so that $\cos(\phi_{j\ell} - \phi_{i\ell}) > 0$, we see that $\mathrm{Re}(\beta_{i\ell}^H \beta_{j\ell}) > 0$ whenever $\beta_{i\ell}^H \beta_{j\ell} \neq 0$. Then

$$\left(1 + \frac{|\mathrm{Im}(\beta_{i\ell}^H \beta_{j\ell})|^2}{|\mathrm{Re}(\beta_{i\ell}^H \beta_{j\ell})|^2} \right)^{1/2} \le \left(1 + \tan^2(\phi_{j\ell} - \phi_{i\ell}) \right)^{1/2} = \frac{1}{\cos(\phi_{j\ell} - \phi_{i\ell})} \le \frac{1}{\cos(2\phi)},$$

where the last step uses (22). Using this in (18) completes the proof. ∎

In Theorem 3(b), we can more generally consider $\beta_{i\ell}$ of the form $\beta_{i\ell} = \omega_{i\ell} e^{\mathbf{i}\phi_{i\ell}}(1 + \mathbf{i}\theta_{i\ell})$, where $\omega_{i\ell} \geq 0$, $\alpha_{i\ell}$ satisfies (19), and

$$|\theta_{j\ell} - \theta_{i\ell}| \leq \sigma |1 + \theta_{i\ell}\theta_{j\ell}| \quad \forall\, i, j, \ell \ \text{ for some } \sigma \geq 0 \text{ with } \tan(2\phi)\sigma < 1. \tag{23}$$

Then the proof of Theorem 3(b) can be extended to show the following upper bound on C given by (18):

$$C \leq \frac{1}{\cos(2\phi)} \cdot \frac{\sqrt{1+\sigma^2}}{1 - \tan(2\phi)\sigma}. \tag{24}$$

However, this generalization is superficial as we can also derive (24) from (20) by rewriting $\beta_{i\ell}$ as

$$\beta_{i\ell} = |\beta_{i\ell}| e^{\mathbf{i}\tilde{\phi}_{i\ell}} \quad \text{with} \quad \tilde{\phi}_{i\ell} = \phi_{i\ell} + \tan^{-1}(\theta_{i\ell}).$$

Then, applying (20) yields $C \geq \cos(2\tilde{\phi})$, where $\tilde{\phi} = \max_{i,j,\ell} |\tilde{\phi}_{i\ell} - \tilde{\phi}_{j\ell}|/2$. Using trigonometric identity, it can be shown that $\cos(2\tilde{\phi})$ equals the right-hand side of (24) with $\sigma = \max_{\{i,j,\ell \ \mid \ \theta_{i\ell}\theta_{j\ell} \neq -1\}} |\theta_{j\ell} - \theta_{i\ell}|/|1 + \theta_{i\ell}\theta_{j\ell}|$.

Notice that Theorem 3(b) implies that if $\phi = 0$, then the SDP relaxation (7) is tight for the quadratically constrained QP (1) with $\mathbb{F} = \mathbb{C}$. Such is the case when all components of h_ℓ, $\ell = 1, \ldots, M$, are real and nonnegative.

4. A convex QP restriction. In this subsection, we consider a convex quadratic programming *restriction* of (2) in the complex case of $\mathbb{F} = \mathbb{C}$ and analyze its approximation bound. Let us write h_ℓ (the channel steering vector) as

$$h_\ell = (\ldots, |h_{j\ell}| e^{\mathbf{i}\phi_{j\ell}}, \ldots)_{j=1,\ldots,n}^T.$$

For any $\bar{\phi}_j \in [0, 2\pi)$, $j = 1, \ldots, n$, and any $\phi \in (0, \pi/2)$, define the four corresponding index subsets

$$\begin{aligned}
J_\ell^1 &:= \{j \mid \phi_{j\ell} \in [\bar{\phi}_j - \phi, \bar{\phi}_j + \phi]\},\\
J_\ell^2 &:= \{j \mid \phi_{j\ell} \in [\bar{\phi}_j - \phi + \pi/2, \bar{\phi}_j + \phi + \pi/2]\},\\
J_\ell^3 &:= \{j \mid \phi_{j\ell} \in [\bar{\phi}_j - \phi + \pi, \bar{\phi}_j + \phi + \pi]\},\\
J_\ell^4 &:= \{j \mid \phi_{j\ell} \in [\bar{\phi}_j - \phi + 3\pi/2, \bar{\phi}_j + \phi + 3\pi/2]\}
\end{aligned}$$

for $\ell = 1, \ldots, M$. The above four subsets are pairwise disjoint if and only if $\phi < \pi/4$ and are collectively exhaustive if and only if $\phi \geq \pi/4$. Choose an index subset J with the property that

$$\text{for each } \ell, \text{ at least one of } J_\ell^1,\ J_\ell^2,\ J_\ell^3,\ J_\ell^4 \text{ contains } J.$$

Of course, $J = \emptyset$ is always allowable, but we should choose J maximally since our approximation bound will depend on the ratio $n/|J|$ (see Theorem 4). Partition the constraint set index $\{1, \ldots, M\}$ into four subsets K^1, K^2, K^3, K^4 such that

$$J \subseteq J_\ell^k \quad \forall\, \ell \in K^k,\ k = 1,2,3,4.$$

Consider the following convex QP restriction of (2) corresponding to K^1, K^2, K^3, K^4:

$$\begin{array}{rcll} \upsilon_{\rm cqp} & := & \min & \|z\|^2 \\ & & \text{s.t.} & \mathrm{Re}(h_\ell^H z) \geq 1 \quad \forall\, \ell \in K^1, \\ & & & -\mathrm{Im}(h_\ell^H z) \geq 1 \quad \forall\, \ell \in K^2, \\ & & & -\mathrm{Re}(h_\ell^H z) \geq 1 \quad \forall\, \ell \in K^3, \\ & & & \mathrm{Im}(h_\ell^H z) \geq 1 \quad \forall\, \ell \in K^4. \end{array} \tag{25}$$

The above problem is a restriction of (2) because for any $z \in \mathbb{C}$,

$$\begin{aligned} |z| &\geq \max\{|\mathrm{Re}(z)|, |\mathrm{Im}(z)|\} \\ &= \max\{\mathrm{Re}(z), \mathrm{Im}(z), -\mathrm{Re}(z), -\mathrm{Im}(z)\}. \end{aligned}$$

If $J \neq \emptyset$ and $(\dots, h_{j\ell}, \dots)_{j\in J} \neq 0$ for $\ell = 1, \dots, M$, then (25) is feasible and hence has an optimal solution. Since (25) is a restriction of (2), $\upsilon_{\rm qp} \leq \upsilon_{\rm cqp}$. We have the following approximation bound.

THEOREM 4. *Suppose that $J \neq \emptyset$ and (25) is feasible. Then*

$$\upsilon_{\rm cqp} \leq \upsilon_{\rm qp}\ \frac{N}{\cos^2\phi} \max_{k=1,\dots,N} \left(\max_{j\in\hat{J}_k} \frac{\bar{\eta}_j}{\underline{\eta}_{\pi_k(j)}} \right)^2,$$

where $N := \lceil n/|J| \rceil$, $\bar{\eta}_j := \max_\ell |h_{j\ell}|$, $\underline{\eta}_j := \min_{\ell | h_{j\ell} \neq 0} |h_{j\ell}|$, $\hat{J}_1, \dots, \hat{J}_N$, is any partition of $\{1, \dots, n\}$ satisfying $|\hat{J}_k| \leq |J|$ for $k = 1, \dots, N$ and π_k is any injective mapping from $\hat{J}_k$ to J.

Proof. By making the substitution

$$z_j^{\rm new} \leftarrow z_j e^{\mathbf{i}\bar{\phi}_j},$$

we can without loss of generality assume that $\bar{\phi}_j = 0$ for all j and ℓ.

Let z^* denote an optimal solution of (2) and write

$$z^* = (\dots, r_j e^{\mathbf{i}\beta_j}, \dots)_{j=1,\dots,n}^T$$

with $r_j \geq 0$. Then, for any ℓ, we have from $|h_{j\ell}| \leq \bar{\eta}_j$ for all j that

$$1 \leq |h_\ell^H z^*| \leq r := \sum_{j=1}^n r_j \bar{\eta}_j.$$

Also, we have

$$\upsilon_{\rm qp} = \|z^*\|^2 = \sum_{j=1}^n r_j^2.$$

Define

$$R_k := \left(\sum_{j\in\hat{J}_k} r_j^2 \right)^{1/2}, \qquad S_k := \sum_{j\in\hat{J}_k} r_j \bar{\eta}_j.$$

Then

$$1 \le r = \sum_{k=1}^{N} S_k, \qquad v_{\rm qp} = \sum_{k=1}^{N} R_k^2.$$

Without loss of generality, assume that $R_1/S_1 = \min_k R_k/S_k$. Then, using the fact that

$$\min_k \frac{|x_k|}{|y_k|} \le \sqrt{N}\frac{\|x\|_2}{\|y\|_1}$$

for any $x, y \in \mathbb{R}^N$ with $y \neq 0$,[4] we see from the above relations that

$$\begin{aligned}\frac{R_1}{S_1} &\le \frac{R_1}{S_1} r \\ &\le \sqrt{N}\frac{\sqrt{v_{\rm qp}}}{r} r \\ &= \sqrt{N}\sqrt{v_{\rm qp}}.\end{aligned}$$

Since $|\hat{J}_1| \le |J|$, there is an injective mapping π from $\hat{J}_1$ to J. Let $\omega := \min_{j\in\hat{J}_1} \underline{\eta}_{\pi(j)}/\bar{\eta}_j$. Define the vector $\bar{z} \in \mathbb{C}^n$ by

$$\bar{z}_j := \begin{cases} r_{\pi^{-1}(j)}/(S_1\omega\cos\phi) & \text{if } j \in \pi(\hat{J}_1); \\ 0 & \text{else.}\end{cases}$$

Then,

$$\|\bar{z}\|^2 = \frac{R_1^2}{S_1^2\omega^2\cos^2\phi} \le \frac{N v_{\rm qp}}{\omega^2\cos^2\phi}.$$

Moreover, for each $\ell \in K^1$, since $\pi(\hat{J}_1) \subseteq J \subseteq J_\ell^1$, we have

$$\begin{aligned}\mathrm{Re}\left(h_\ell^H \bar{z}\right) &= \mathrm{Re}\left(\sum_{j\in\pi(\hat{J}_1)} h_{j\ell}^H \bar{z}_j\right) \\ &= \frac{1}{S_1\omega\cos\phi}\mathrm{Re}\left(\sum_{j\in\pi(\hat{J}_1)} r_{\pi^{-1}(j)}|h_{j\ell}|e^{-\mathbf{i}\phi_{j\ell}}\right) \\ &= \frac{1}{S_1\omega\cos\phi}\sum_{j\in\pi(\hat{J}_1)} r_{\pi^{-1}(j)}|h_{j\ell}|\cos\phi_{j\ell} \\ &\ge \frac{1}{S_1\omega\cos\phi}\sum_{j\in\pi(\hat{J}_1)} r_{\pi^{-1}(j)}\underline{\eta}_j\cos\phi \\ &= \frac{1}{S_1\omega}\sum_{j\in\hat{J}_1} r_j\bar{\eta}_j\,\frac{\underline{\eta}_{\pi(j)}}{\bar{\eta}_j} \\ &\ge \frac{1}{S_1\omega}\sum_{j\in\hat{J}_1} r_j\bar{\eta}_j\cdot\min_{j\in\hat{J}_1}\frac{\underline{\eta}_{\pi(j)}}{\bar{\eta}_j} \\ &= 1,\end{aligned}$$

[4] *Proof.* Suppose the contrary, so that for some $x, y \in \mathbb{R}^N$ with $y \neq 0$, we have $|x_k|/|y_k| > \sqrt{N}\|x\|_2/\|y\|_1$ for all k. Then, multiplying both sides by $|y_k|$ and summing over k yields $\|x\|_1 > \sqrt{N}\|x\|_2$, contradicting properties of 1- and 2-norms.

where the first inequality uses $|h_{j\ell}| \geq \underline{\eta}_j$ and $\phi_{j\ell} \in [-\phi, \phi]$ for $j \in J_\ell^1$. Since $\bar{z}_j = 0$ for $j \notin J_\ell^1$, this shows that $\bar{z}$ satisfies the first set of constraints in (25). A similar reasoning shows that $\bar{z}$ satisfies the remaining three sets of constraints in (25). □

Notice that the $\bar{z}$ constructed in the proof of Theorem 4 is feasible for the further restriction of (25) whereby $z_j = 0$ for all $j \notin J$. This further restricted problem has the same (worst-case) approximation bound specified in Theorem 4.

Let us compare the two approximation bounds in Theorems 3 and 4. First, the required assumptions are different. On the one hand, the bound in Theorem 3 does not depend on $|h_{j\ell}|$, while the bound in Theorem 4 does. On the other hand, Theorem 3 requires that the bounded angular spread

$$|\phi_{j\ell} - \phi_{i\ell}| \leq 2\phi \quad \forall\, j, \ell \tag{26}$$

for some $\phi < \pi/4$, while Theorem 4 allows $\phi < \pi/2$ and requires only the condition (26) for all $1 \leq \ell \leq M$ and $j \in J$, where J is a preselected index set. Thus, the bounded angular spread condition required in Theorem 3 corresponds exactly to $|J| = n$. Thus, the assumptions required in the two theorems do not imply one another. Second, the two performance ratios are also different. Naturally, the final performance ratio in Theorem 4 depends on the choice of J through the ratio $|J|/n$, so a large J is preferred. In the event that the assumptions of both theorems are satisfied and let us assume for simplicity that $\bar{\eta}_j = \underline{\eta}_j$ for all j, then $|J| = n$ and $\phi < \pi/4$, in which case Theorem 4 gives a performance ratio of $1/\cos^2\phi$ while Theorem 3 gives $1/\cos(2\phi)$. Since $\cos(2\phi) = \cos^2\phi - \sin^2\phi \leq \cos^2\phi$, we have $1/\cos(2\phi) \geq 1/\cos^2\phi$, showing that Theorem 4 gives a tighter approximation bound. However, this does not mean Theorem 4 is stronger than Theorem 3 since the two theorems hold under different assumptions in general.

We can specialize Theorem 4 to a typical situation in transmit beamforming. Consider a uniform linear transmit antenna array consisting of n elements, and let us assume that the M receivers are in a sector area from the far field and the propagation is line-of-sight. By reciprocity, each steering vector h_ℓ will be Vandermonde with generator $e^{-\mathbf{i}2\pi\frac{d}{\lambda}\sin\theta_\ell}$ (see, e.g., [10]), where d is the interantenna spacing, λ is the wavelength, and θ_ℓ is the angle of arrival of the ℓth receiving antenna. In a sector of approximately 60 degrees about the array broadside, we will have $|\theta_\ell| \leq \pi/3$. Suppose that $d/\lambda = 1/2$. Then the steering vector corresponding to the ℓth receiving antenna will have the form

$$h_\ell = (\ldots, e^{-\mathbf{i}(j-1)\pi\sin\theta_\ell}, \ldots)_{j=1,\ldots,n}^T.$$

In this case, we have that $\phi_{j\ell} = (j-1)\pi\sin\theta_\ell$ and $|h_{j\ell}| = 1$ for all j and ℓ. We can take, e.g.,

$$\bar{\phi}_j = 0, \quad \phi = \bar{j}\pi \max_\ell |\sin\theta_\ell|, \quad J = \{1, \ldots, \bar{j}+1\},$$

where $\bar{j} := \lfloor 1/\max_\ell |\sin\theta_\ell| \rfloor$. Thus, the assumptions of Theorem 4 are satisfied. Moreover, since $|\theta_\ell| \leq \pi/3$ for all ℓ, it follows that $|J| = \bar{j} + 1 \geq 2$. If n is not large, say, $n \leq 8$, then Theorem 4 gives a performance ratio of $n/(|J|\cos^2\phi) \leq 16$.

More generally, if we can choose the partition $\hat{J}_1, \ldots, \hat{J}_N$ and the mapping π_k in Theorem 4 such that

$$(\dots,\bar{\eta}_j,\dots)_{j\in\hat{J}_k} = (\dots,\underline{\eta}_{\pi_k(j)},\dots)_{j\in J} \quad \forall\, k,$$

then the performance ratio in Theorem 4 simplifies to $N/\cos^2\phi$. In particular, this holds when $|h_{j\ell}| = \eta > 0$ for all j and ℓ or when $J = \{1,\dots,n\}$ (so that $N = 1$) and $|h_{j\ell}|$ is independent of ℓ for all j, and, more generally, when the channel coefficients periodically repeat their magnitudes. In general, we should choose the partition $\hat{J}_1,\dots,\hat{J}_N$ and the mapping π_k to make the performance ratio in Theorem 4 small. For example, if $J = \hat{J}_1 = \{1,2\}$ and $\bar{\eta}_1 = 100$, $\bar{\eta}_2 = 10$, $\underline{\eta}_1 = 1$, $\underline{\eta}_2 = 10$, then $\pi_1(1) = 2, \pi_1(2) = 1$ is the better choice.

5. Homogeneous QP in maximization form. Let us now consider the following complex norm maximization problem with convex homogeneous quadratic constraints:

$$\begin{array}{rll} v_{\rm qp} := \max & \|z\|^2 & \\ \text{s.t.} & \displaystyle\sum_{\ell\in\mathcal{I}_i} |h_\ell^H z|^2 \le 1, & i = 1,\dots,m, \\ & z\in\mathbb{C}^n, & \end{array} \tag{27}$$

where $h_\ell \in \mathbb{C}^n$.

To motivate this problem, consider the problem of designing an intercept beamformer[5] capable of suppressing signals impinging on the receiving antenna array from irrelevant or hostile emitters, e.g., jammers, whose steering vectors (spatial signatures, or "footprints") have been previously estimated, while achieving as high gain as possible for all other transmissions. The jammer suppression capability is captured in the constraints of (27), and $|\mathcal{I}_i| > 1$ covers the case where a jammer employs more than one transmit antenna. The maximization of the objective $\|z\|^2$ can be motivated as follows. In intercept applications, the steering vector of the emitter of interest, h, is a priori unknown and is naturally modeled as random. A pertinent optimization objective is then the average beamformer output power, measured by $E[|h^H z|^2]$. Under the assumption that the entries of h are uncorrelated and have equal average power, it follows that $E[|h^H z|^2]$ is proportional to $\|z\|^2$, which is often referred to as the beamformer's *white noise gain*.

Similar to (1), we let

$$H_i := \sum_{\ell\in\mathcal{I}_i}^{m} h_\ell h_\ell^H$$

and consider the natural SDP relaxation of (27):

$$\begin{array}{rll} v_{\rm sdp} := \max & \mathrm{Tr}(Z) & \\ \text{s.t.} & \mathrm{Tr}(H_i Z) \le 1, & i = 1,\dots,m, \\ & Z \succeq 0, & Z \text{ is complex and Hermitian.} \end{array} \tag{28}$$

We are interested in lower bounds for the relaxation performance of the form

$$v_{\rm qp} \ge C\; v_{\rm sdp},$$

where $0 < C \le 1$. It is easily checked that (28) has an optimal solution.

Let Z^* be an optimal solution of (28). We will analyze the performance of the SDP relaxation using the following randomization procedure:

[5]Note that here we are talking about a receive beamformer, as opposed to our earlier motivating discussion of transmit beamformer design.

1. Generate a random vector $\xi \in \mathbb{C}^n$ from the *complex-valued* normal distribution $N_c(0, Z^*)$.
2. Let $z^*(\xi) = \xi / \max_{1 \le i \le m} \sqrt{\xi^H H_i \xi}$.

First, we need the following lemma analogous to Lemmas 1 and 3.

LEMMA 5. *Let $H \in \mathbb{C}^{n \times n}$, $Z \in \mathbb{C}^{n \times n}$ be two Hermitian positive semidefinite matrices (i.e., $H \succeq 0$, $Z \succeq 0$). Suppose $\xi \in \mathbb{C}^n$ is a random vector generated from the complex-valued normal distribution $N_c(0, Z)$. Then, for any $\gamma > 0$,*

$$\text{Prob}\left(\xi^H H \xi > \gamma E(\xi^H H \xi)\right) \le \bar{r}\ e^{-\gamma}, \tag{29}$$

where $\bar{r} := \min\{\text{rank}(H), \text{rank}(Z)\}$.

Proof. If $H = 0$, then (29) is trivially true. Suppose $H \neq 0$. Then, as in the proof of Lemma 1, we have

$$\text{Prob}\left(\xi^H H \xi > \gamma E(\xi^H H \xi)\right) = \text{Prob}\left(\sum_{i=1}^{\bar{r}} \bar{\lambda}_i |\bar{\xi}_i|^2 > \gamma\right),$$

where $\bar{\lambda}_1 \ge \bar{\lambda}_2 \ge \cdots \ge \bar{\lambda}_{\bar{r}} \ge 0$ satisfy $\bar{\lambda}_1 + \cdots + \bar{\lambda}_{\bar{r}} = 1$ and each $\bar{\xi}_i \in \mathbb{C}$ has the complex-valued normal distribution $N_c(0, 1)$. Then

$$\begin{aligned}
\text{Prob}\left(\xi^H H \xi > \gamma E(\xi^H H \xi)\right) &\le \text{Prob}\left(|\bar{\xi}_1|^2 > \gamma \text{ or } |\bar{\xi}_2|^2 > \gamma \text{ or } \cdots \text{ or } |\bar{\xi}_{\bar{r}}|^2 > \gamma\right) \\
&\le \sum_{i=1}^{\bar{r}} \text{Prob}\left(|\bar{\xi}_i|^2 > \gamma\right) \\
&= \bar{r}\ e^{-\gamma},
\end{aligned}$$

where the last step uses (14). □

THEOREM 5. *For the complex QP* (27) *and its SDP relaxation* (28)*, we have $v_{\text{sdp}} = v_{\text{qp}}$ if $m \le 3$ and otherwise*

$$v_{\text{qp}} \ge \frac{1}{4 \ln(100K)} v_{\text{sdp}},$$

where $K := \sum_{i=1}^m \min\{\text{rank}(H_i), \sqrt{m}\}$.

Proof. By applying a suitable rank reduction procedure if necessary, we can assume that the rank r of the optimal SDP solution Z^* satisfies $r = 1$ if $m \le 3$ and $r \le \sqrt{m}$ if $m \ge 4$; see [9, section 5]. Thus, if $m \le 3$, then $Z^* = z^*(z^*)^H$ for some $z^* \in \mathbb{C}^n$ and it is readily seen that z^* is an optimal solution of (27), so that $v_{\text{sdp}} = v_{\text{qp}}$. Otherwise, we apply the randomization procedure to Z^*. By using Lemma 5, we have, for any $\gamma > 0$ and $\mu > 0$,

$$\begin{aligned}
&\text{Prob}\left(\max_{1 \le i \le m} \xi^H H_i \xi \le \gamma,\ \|\xi\|^2 \ge \mu \text{Tr}(Z^*)\right) \\
&\ge 1 - \sum_{i=1}^m \text{Prob}\left(\xi^H H_i \xi > \gamma E(\xi^H H_i \xi)\right) - \text{Prob}\left(\|\xi\|^2 < \mu \text{Tr}(Z^*)\right) \\
&\ge 1 - K e^{-\gamma} - \text{Prob}\left(\|\xi\|^2 < \mu \text{Tr}(Z^*)\right), \qquad (30)
\end{aligned}$$

where the last step uses $r \le \sqrt{m}$.

Let

$$\eta_j := \begin{cases} |\xi_j|^2/Z^*_{jj} & \text{if } Z^*_{jj} > 0; \\ 0 & \text{if } Z^*_{jj} = 0, \end{cases} \quad j = 1, \ldots, n.$$

For simplicity, let us assume that $Z^*_{jj} > 0$ for all $j = 1, \ldots, n$. Since $\xi_j \sim N_c(0, Z^*_{jj})$, as we discussed in subsection 3.2, $|\xi_j|$ follows a Weibull distribution with variance Z^*_{jj} (see (14)), and therefore

$$\text{Prob}\,(\eta_j \leq t) = 1 - e^{-t} \quad \forall\, t \in [0, \infty).$$

Hence,

$$E(\eta_j) = \int_0^\infty te^{-t}dt = 1, \quad E(\eta_j^2) = \int_0^\infty t^2 e^{-t}dt = 2, \quad \text{Var}(\eta_j) = 1.$$

Moreover,

$$E(|\eta_j - E(\eta_j)|) = \int_0^1 (1-t)e^{-t}dt + \int_1^\infty (t-1)e^{-t}dt = \frac{2}{e}.$$

Let us denote $\lambda_j = Z^*_{jj}/\text{Tr}(Z^*)$, $j = 1, \ldots, n$, and $\eta := \sum_{j=1}^n \lambda_j \eta_j$. We have $E(\eta) = 1$ and

$$E(|\eta - E(\eta)|) = E\left(\left|\sum_{j=1}^n \lambda_j(\eta_j - E(\eta_j))\right|\right) \leq \sum_{j=1}^n \lambda_j E(|\eta_j - E(\eta_j)|) = \frac{2}{e}.$$

Since, by Markov's inequality,

$$\text{Prob}\,(|\eta - E(\eta)| > \alpha) \leq \frac{E(|\eta - E(\eta)|)}{\alpha} \leq \frac{2}{\alpha e} \quad \forall\, \alpha > 0,$$

we have

$$\begin{aligned} \text{Prob}\left(\|\xi\|^2 < \mu \text{Tr}(Z^*)\right) &= \text{Prob}\,(\eta < \mu) \\ &\leq \text{Prob}\,(|\eta - E(\eta)| > 1 - \mu) \\ &\leq \frac{2}{e(1-\mu)} \quad \forall\, \mu \in (0,1). \end{aligned}$$

Substituting the above inequality into (30), we obtain

$$\text{Prob}\left(\max_{1\leq i\leq m} \xi^H H_i \xi \leq \gamma,\ \|\xi\|^2 \geq \mu \text{Tr}(Z^*)\right) > 1 - \text{Ke}^{-\gamma} - \frac{2}{e(1-\mu)} \quad \forall\, \mu \in (0,1).$$

Setting $\mu = 1/4$ and $\gamma = \ln(100K)$ yields a positive right-hand side of 0.00898.., which then proves the desired bound. □

The above proof technique also applies to the real case, i.e., $h_\ell \in \mathbb{R}^n$ and $z \in \mathbb{R}^n$. The main difference is that $\xi \sim N(0, Z^*)$, so that $|\bar{\xi}_i|^2$ in the proof of Lemma 5 and η_j in the proof of Theorem 5 both follow a χ^2 distribution with one degree of freedom. Then

$$\text{Prob}\left(|\bar{\xi}_i|^2 > \gamma\right) = \int_{\sqrt{\gamma}}^\infty \frac{e^{-t^2/2}}{\sqrt{2\pi}}dt \leq \int_{\sqrt{\gamma}}^\infty \frac{e^{-\gamma t/2}}{\sqrt{2\pi}}dt = \sqrt{\frac{2}{\pi\gamma}}e^{-\gamma/2} \quad \forall\, \gamma > 0,$$

$E(\eta_j) = 1$, and

$$\begin{aligned} E|\eta_j - E(\eta_j)| &= \int_0^\infty \frac{e^{-t/2}}{\sqrt{2\pi t}}|t-1|dt \\ &= \frac{1}{\sqrt{2\pi}}\int_0^1 \frac{e^{-t/2}}{\sqrt{t}}dt - \frac{1}{\sqrt{2\pi}}\int_0^1 \sqrt{t}e^{-t/2}dt \\ &\quad + \frac{1}{\sqrt{2\pi}}\int_1^\infty \sqrt{t}e^{-t/2}dt - \frac{1}{\sqrt{2\pi}}\int_1^\infty \frac{e^{-t/2}}{\sqrt{t}}dt \\ &= \frac{4}{\sqrt{2\pi e}} < 0.968, \end{aligned}$$

where in the last step we used integration by parts on the first and the fourth terms. This yields the analogous bound that for any $\gamma \geq 1$ and $\mu \in (0,1)$,

$$\begin{aligned} \text{Prob}\left(\max_{1\leq i\leq m} \xi^T H_i \xi \leq \gamma,\ \|\xi\|^2 \geq \mu \text{Tr}(Z^*)\right) &> 1 - K\sqrt{\frac{2}{\pi\gamma}}e^{-\gamma/2} - \frac{0.968}{1-\mu} \\ &> 1 - Ke^{-\gamma/2} - \frac{0.968}{1-\mu}, \end{aligned}$$

where $K := \sum_{i=1}^m \min\{\text{rank}(H_i), \sqrt{2m}\}$. Setting $\mu = 0.01$ and $\gamma = 2\ln(50K)$ yields a positive right-hand side of 0.0022... This in turn shows that $v_{\text{sdp}} = v_{\text{qp}}$ if $m \leq 2$ (see the proof of Theorem 1) and otherwise

$$v_{\text{qp}} \geq \frac{1}{200\ln(50K)} v_{\text{sdp}}.$$

We note that, in the real case, a sharper bound of

$$v_{\text{qp}} \geq \frac{1}{2\ln(2m\mu)} v_{\text{sdp}},$$

where $\mu := \min\{m, \max_i \text{rank}(H_i)\}$, was shown by Nemirovski et al. [14] (also see [13, Theorem 4.7]), although the above proof seems simpler. Also, an example in [14] shows that the $O(1/\ln m)$ bound is tight (up to a constant factor) in the worst case. This example readily extends to the complex case by identifying $\mathbb{C}^n$ with $\mathbb{R}^{2n}$ and observing that $|h_\ell^H z| \geq |\text{Re}(h_\ell)^T \text{Re}(z) + \text{Im}(h_\ell)^T \text{Im}(z)|$ for any $h_\ell, z \in \mathbb{C}^n$. Thus, in the complex case, the $O(1/\ln m)$ bound is also tight (up to a constant factor).

6. Discussion. In this paper, we have analyzed the worst-case performance of SDP relaxation and convex restriction for a class of NP-hard quadratic optimization problems with homogeneous quadratic constraints. Our analysis is motivated by important emerging applications in transmit beamforming for physical layer multicasting and sensor localization in wireless sensor networks. Our generalization (1) of the basic problem in [20] is useful, for it shows that the same convex approximation approaches and bounds hold in the case where each multicast receiver is equipped with multiple antennas. This scenario is becoming more pertinent with the emergence of small and cheap multiantenna mobile terminals. Furthermore, our consideration of the related homogeneous QP maximization problem has direct application to the design of jam-resilient intercept beamformers. In addition to these timely topics, more traditional signal processing design problems can be cast in the same mathematical framework; see [20] for further discussions.

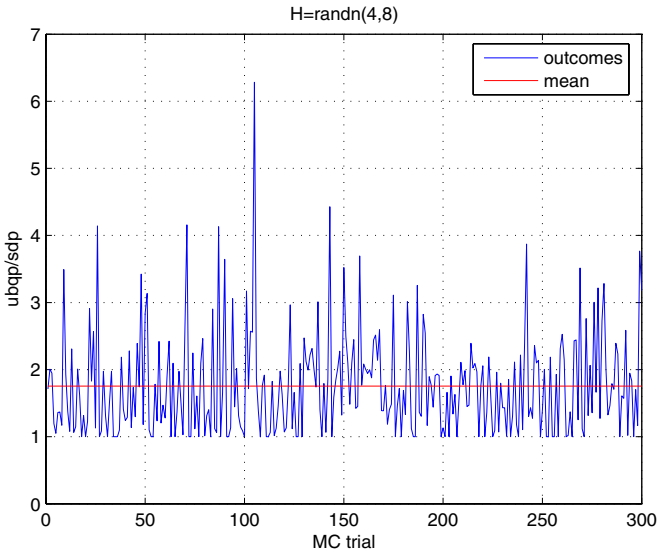


FIG. 1. *Upper bound on* $\frac{v_{\rm qp}}{v_{\rm sdp}}$ *for* $m = 8$, $n = 4$, 300 *realizations of real Gaussian i.i.d. steering vector entries, solution constrained to be real.*

While theoretical worst-case analysis is very useful, empirical analysis of the ratio $\frac{v_{\rm qp}}{v_{\rm sdp}}$ through simulations with randomly generated steering vectors $\{h_\ell\}$ is often equally important. In the context of transmit beamforming for multicasting [20] for the case $|\mathcal{I}_i| = 1$ for all i (single receiving antenna per subscriber node), simulations have provided the following insights:

- For moderate values of m, n (e.g., $m = 24$, $n = 8$), and independent and identically distributed (i.i.d.) complex-valued circular Gaussian (i.i.d. Rayleigh) entries of the steering vectors $\{h_\ell\}$, the average value of $\frac{v_{\rm qp}}{v_{\rm sdp}}$ is under 3—much lower than the worst-case value predicted by our analysis.
- In all generated instances where all steering vectors have positive real and imaginary parts, the ratio $\frac{v_{\rm qp}}{v_{\rm sdp}}$ equals one (with error below 10^{-8}). This is better than what our worst-case analysis predicts for limited phase spread (see Theorem 3).
- In experiments with measured VDSL channel data, for which the steering vectors follow a correlated log-normal distribution, $\frac{v_{\rm qp}}{v_{\rm sdp}} = 1$ in over 50% of instances.
- Our analysis shows that the worst-case performance ratio $\frac{v_{\rm qp}}{v_{\rm sdp}}$ is smaller in the complex case than in the real case ($O(m)$ versus $O(m^2)$). Moreover, this remains true with high probability when $v_{\rm qp}$ is replaced by its upper bound

$$v_{\rm ubqp} := \min_{k=1,\ldots,N} \|z^*(\xi^k)\|^2,$$

where $\xi^1, \ldots, \xi^N$ are generated by N independent trials of the randomization procedure (see subsections 3.1 and 3.2) and N is taken sufficiently large. In our simulation, we used $N = 30nm$. Figure 1 shows our simulation results

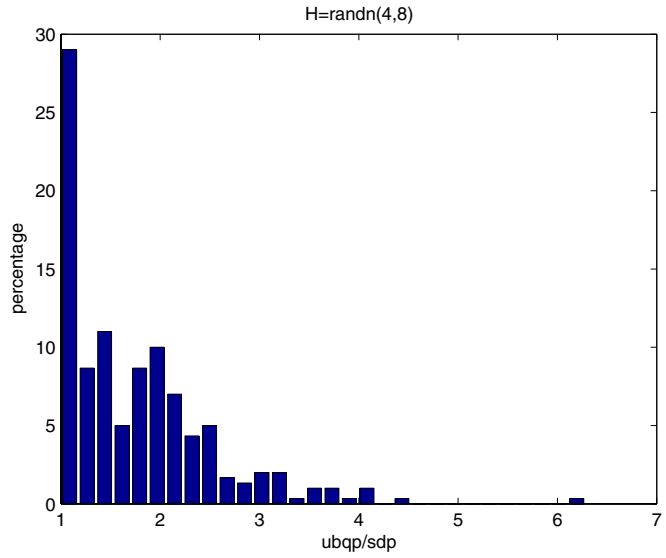


FIG. 2. *Histogram of the outcomes in Figure* 1.

for the real Gaussian case.[6] It plots $\frac{v_{\text{ubqp}}}{v_{\text{sdp}}}$ for 300 independent realizations of i.i.d. real-valued Gaussian steering vector entries for $m = 8$, $n = 4$. Figure 2 plots the corresponding histogram. Figures 3 and 4 show the corresponding results for i.i.d. complex-valued circular Gaussian steering vector entries.[7] Both the mean and the maximum of the upper bound $\frac{v_{\text{ubqp}}}{v_{\text{sdp}}}$ are lower in the complex case. The simulations indicate that SDP approximation is better in the complex case not only in the worst case but also on average.

The above empirical (worst-case and average-case) analysis complements our theoretical worst-case analysis of the performance of SDP relaxation for the class of problems considered herein.

Finally, we remark that our worst-case analysis of SDP performance is based on the assumption that the homogeneous quadratic constraints are concave (see (1)). Can we extend this analysis to general homogeneous quadratic constraints? The following example in $\mathbb{R}^2$ suggests that this is not possible.

Example 3. For any $L > 0$, consider the quadratic optimization problem with homogeneous quadratic constraints:

$$\begin{array}{ll} \min & \|z\|^2 \\ \text{s.t.} & z_2^2 \geq 1, \quad z_1^2 - L z_1 z_2 \geq 1, \quad z_1^2 + L z_1 z_2 \geq 1, \\ & z \in \mathbb{R}^2. \end{array} \tag{31}$$

The last two constraints imply $z_1^2 \geq L|z_1||z_2| + 1$ which, together with the first constraint $z_2^2 \geq 1$, yield $z_1^2 \geq L|z_1| + 1$ or, equivalently, $|z_1| \geq (L + \sqrt{L^2+4})/2$. So the optimal value of (31) is at least $1 + (L + \sqrt{L^2+4})^2/4$ (and in fact is equal to this).

[6]Here the SDP solution is constrained to be real-valued, and real Gaussian randomization is used.

[7]Here the SDP solutions are complex-valued, and complex Gaussian randomization is used.

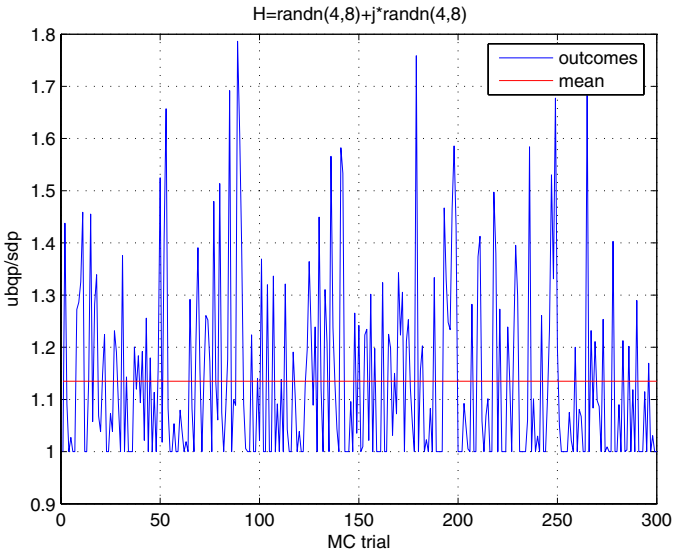


FIG. 3. *Upper bound on* $\frac{v_{\rm qp}}{v_{\rm sdp}}$ *for* $m = 8$, $n = 4$, 300 *realizations of complex Gaussian i.i.d. steering vector entries.*

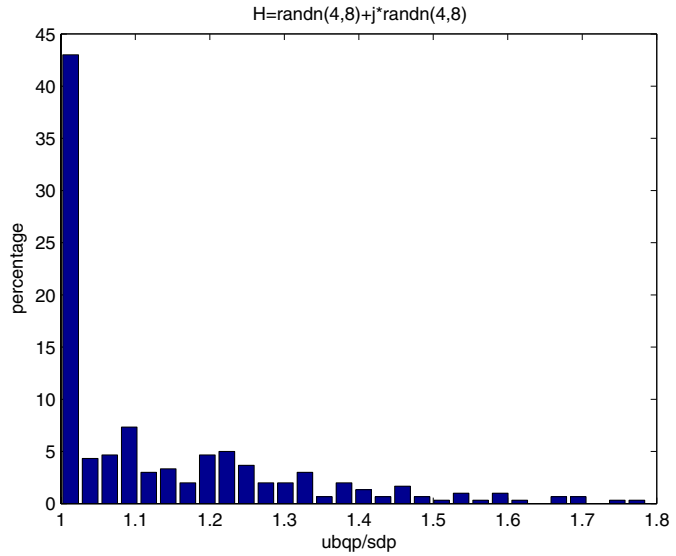


FIG. 4. *Histogram of the outcomes in Figure* 3.

The natural SDP relaxation of (31) is

$$\begin{array}{ll} \min & Z_{11} + Z_{22} \\ \text{s.t.} & Z_{22} \geq 1, \quad Z_{11} - LZ_{12} \geq 1, \quad Z_{11} + LZ_{12} \geq 1, \\ & Z \succeq 0. \end{array}$$

Clearly, $Z = I_2$ is a feasible solution (and, in fact, an optimal solution) of this SDP, with an objective value of 2. Therefore, the SDP performance ratio for this example is at least $1/2 + (L + \sqrt{L^2+4})^2/8$, which can be arbitrarily large.

REFERENCES

[1] N. Alon and A. Naor, *Approximating the Cut-Norm via Grothendieck's inequality,* Proceedings of the 36th Annual ACM Symposium on Theory of Computing, Chicago, IL, 2004, pp. 72–80.

[2] H. H. Andersen, M. Højbjerre, D. Sørensen, and P. S. Eriksen, *Linear and Graphical Models for the Multivariate Complex Normal Distribution,* Lecture Notes in Statist. 101, Springer-Verlag, New York, 1995.

[3] A. Ben-Tal and A. Nemirovski, *On tractable approximations of uncertain linear matrix inequalities affected by interval uncertainty*, SIAM J. Optim., 12 (2002), pp. 811–833.

[4] A. Ben-Tal, A. Nemirovski, and C. Roos, *Extended matrix cube theorems with applications to μ-theory in control*, Math. Oper. Res., 28 (2003), pp. 497–523.

[5] D. Bertsimas and Y. Ye, *Semidefinite relaxations, multivarite normal distribution, and order statistics*, in Handbook of Combinatorial Optim., Vol. 3, D. Z. Du and P. M. Pardalos, eds., Kluwer Academic Publishers, Boston, 1998, pp. 1–19.

[6] B. Biswas and Y. Ye, *Semidefinite Programming for Ad Hoc Wireless Sensor Network Localization*, Technical report, Department of Electrical Engineering, Stanford University, Stanford, CA, Sept. 2003; also available online from http://www.stanford.edu/~yyye/.

[7] M. X. Goemans and D. P. Williamson, *Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming*, J. ACM, 42 (1995), pp. 1115–1145.

[8] M. X. Goemans and D. P. Williamson, *Approximation algorithms for MAX-3-CUT and other problems via complex semidefinite programming*, J. Comput System Sci., 68 (2004), pp. 442–470.

[9] Y. Huang and S. Zhang, *Complex matrix decomposition and quadratic programming*, Math. Oper. Res., to appear.

[10] D. H. Johnson and D. E. Dugeon, *Array Signal Processing: Concepts and Techniques,* Simon & Schuster, New York, 1992.

[11] L. Lovász and A. Schrijver, *Cones of matrices and set-functions and* 0-1 *optimization*, SIAM J. Optim., 1 (1991), pp. 166–190.

[12] S. Mahajan and H. Ramesh, *Derandomizing approximation algorithms based on semidefinite programming*, SIAM J. Comput., 28 (1999), pp. 1641–1663.

[13] A. Megretski, *Relaxations of quadratic programs in operator theory and system analysis*, Oper. Theory Adv. Appl., 129 (2001), pp. 365–392.

[14] A. Nemirovski, C. Roos, and T. Terlaky, *On maximization of quadratic form over intersection of ellipsoids with common center*, Math. Program., 86 (1999), pp. 463–473.

[15] Y. Nesterov, *Semidefinite relaxation and nonconvex quadratic optimization*, Optim. Methods Softw., 9 (1998), pp. 141–160.

[16] Y. Nesterov, H. Wolkowicz, and Y. Ye, *Semidefinite programming relaxations of nonconvex quadratic optimization*, in Handbook of Semidefinite Programming, H. Wolkowicz, R. Saigal, and L. Vandenberghe, eds., Kluwer, Boston, 2000, pp. 360–419.

[17] G. Pataki, *On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues*, Math. Oper. Res., 23 (1998), pp. 339–358.

[18] G. Pataki, ed., *Computational semidefinite and second order cone programming: The state of the art*, Math. Program., 95 (2003).

[19] N. Z. Shor, *Quadratic optimization problems*, Soviet J. Comput. Systems Sci., 25 (1987), pp. 1–11.

[20] N. D. Sidiropoulos, T. N. Davidson, and Z.-Q. Luo, *Transmit beamforming for physical layer multicasting*, IEEE Trans. Signal Process., 54 (2006), pp. 2239–2251.

[21] A. So, J. Zhang, and Y. Ye, *On Approximating Complex Quadratic Optimization Problems via Semidefinite Programming Relaxations*, in Proceedings of the 11th Conference in Integer Programming and Combinatorial Optimization, Lecture Notes in Comput. Sci. 3509, M. Junger and V. Kaibel, eds., Springer-Verlag, Berlin, 2005, pp. 125–135.

[22] P. Tseng, *Further results on approximating nonconvex quadratic optimization by semidefinite programming relaxation*, SIAM J. Optim., 14 (2003), pp. 268–283.

[23] Y. Ye, *Approximating quadratic programming with bound and quadratic constraints*, Math. Program., 84 (1999), pp. 219–226.

[24] Y. Ye, *Approximating global quadratic optimization with convex quadratic constraints*, J. Global Optim., 15 (1999), pp. 1–17.
[25] S. Zhang, *Quadratic maximization and semidefinite relaxation*, Math. Program., 87 (2000), pp. 453–465.
[26] S. Zhang and Y. Huang, *Complex quadratic optimization and semidefinite programming*, SIAM J. Optim., 16 (2006), pp. 871–890.

SIAM J. OPTIM.
Vol. 18, No. 1, pp. 29–51

A CONVERGENT INCREMENTAL GRADIENT METHOD WITH A CONSTANT STEP SIZE*

DORON BLATT†, ALFRED O. HERO‡, AND HILLEL GAUCHMAN§

Abstract. An incremental aggregated gradient method for minimizing a sum of continuously differentiable functions is presented. The method requires a single gradient evaluation per iteration and uses a constant step size. For the case that the gradient is bounded and Lipschitz continuous, we show that the method visits infinitely often regions in which the gradient is small. Under certain unimodality assumptions, global convergence is established. In the quadratic case, a global linear rate of convergence is shown. The method is applied to distributed optimization problems arising in wireless sensor networks, and numerical experiments compare the new method with other incremental gradient methods.

Key words. incremental gradient method, convergence analysis, sensor networks, neural networks, logistic regression, boosting

AMS subject classifications. 90C30, 49M37, 65K05

DOI. 10.1137/040615961

1. Introduction.

Consider the unconstrained optimization problem

$$\text{minimize} \qquad f(x) = \sum_{l=1}^{L} f_l(x), \quad x \in \mathbb{R}^p, \tag{1.1}$$

where $\mathbb{R}^p$ is the p-dimensional Euclidean space, and $f_l : \mathbb{R}^p \to \mathbb{R}$ are continuously differentiable scalar functions on $\mathbb{R}^p$. Our interest in this problem stems from optimization problems arising in wireless sensor networks (see, e.g., [9, 33, 36, 37, 38]), in which $f_l(x)$ corresponds to the data collected by the lth sensor in the network. This problem also arises in neural network training, in which $f_l(x)$ corresponds to the lth training data set (see, e.g., [7, 17, 18, 27, 28, 26]).

The iterative method proposed and analyzed in this paper for solving (1.1), which we call the *incremental aggregated gradient* (IAG) method, generates a sequence $\{x^k\}_{k\geq 1}$ as follows. Given L arbitrary initial points $x^1, x^2, \ldots, x^L$, an aggregated gradient, denoted by d^L, is defined as $\sum_{l=1}^{L} \nabla f_l(x^l)$. Possible initializations are discussed in section 3. For $k \geq L$,

$$x^{k+1} = x^k - \mu \frac{1}{L} d^k, \tag{1.2}$$

$$d^{k+1} = d^k - \nabla f_{(k+1)_L}(x^{k+1-L}) + \nabla f_{(k+1)_L}(x^{k+1}), \tag{1.3}$$

*Received by the editors September 29, 2005; accepted for publication (in revised form) July 17, 2006; published electronically February 2, 2007.

http://www.siam.org/journals/siopt/18-1/61596.html

†DRW Trading Group, 10 South Riverside Plaza, 21st Floor, Chicago, IL 60606 (dblatt@drwholdings.com). This work was conducted while this author was at the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109. This author's work was supported in part by NIH/NCI grant 1P01 CA87634, by DARPA-MURI grant ARO DAAD 19-02-1-0262, and by NSF contract CCR-0325571.

‡Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109 (hero@eecs.umich.edu). This author's work was supported in part by NIH/NCI grant 1P01 CA87634, by DARPA-MURI grant ARO DAAD 19-02-1-0262, and by NSF contract CCR-0325571.

§Department of Mathematics and Computer Science, Eastern Illinois University, Charleston, IL 61920 (cfhvg@eiu.edu).

where μ is a positive constant step size chosen small enough to ensure convergence, $(k)_L$ denotes k modulo L with representative class $\{1, 2, \ldots, L\}$, and the factor $1/L$ is explicitly included to make the approximate descent direction $\frac{1}{L}d^k$ comparable in magnitude to the one used in the standard incremental gradient method to be discussed below. Thus, at every iteration a new point x^{k+1} is generated according to the direction of the aggregated gradient d^k. Then only one of the gradient summands $\nabla f_{(k+1)_L}(x^{k+1})$ is computed to replace the previously computed $\nabla f_{(k+1)_L}(x^{k+1-L})$. Note that for $k \geq L$ the IAG iteration (1.2)–(1.3) is equivalent to

$$x^{k+1} = x^k - \mu \frac{1}{L} \sum_{l=0}^{L-1} \nabla f_{(k-l)_L}(x^{k-l}). \tag{1.4}$$

The IAG method is related to the large class of incremental gradient methods that has been studied extensively in the literature [8, 17, 18, 19, 21, 25, 26, 28, 44] (see also [22, 32] and the references therein for incremental subgradient methods for nondifferentiable convex optimization). The standard incremental gradient method updates x^k according to

$$x^{k+1} = x^k - \mu(k) \nabla f_{(k)_L}(x^k), \tag{1.5}$$

where $\mu(k)$ is a positive step size, possibly depending on k. Therefore, it is seen that the principal difference between the two methods is that the standard incremental gradient method uses only one of the components in order to generate an approximate descent direction, whereas the IAG method uses the average of the L previously computed gradients. This property leads to convergence of the IAG method for fixed and sufficiently small positive step size μ. This is in contrast to the standard incremental gradient method, whose convergence requires that the step size sequence $\mu(k)$ converge to zero.

Incremental gradient methods can be motivated by the observation that when the iterates are far from the eventual limit, the evaluation of a single gradient component is sufficient for generating an approximate descent direction. Hence, these methods lead to a significant reduction in the amount of required computations per iteration (see, e.g., [6, sect. 1.5.2] and the discussion in [5]). The drawback of these methods, when using a constant step size, is that the iterates converge to a limit cycle and oscillate around a stationary point [25], unless restrictions of the type $\nabla f_l(x) = 0$, $l = 1, \ldots, L$, whenever $\nabla f(x) = 0$ are imposed [44]. Convergence for a diminishing step size has been established by a number of authors under different conditions [8, 17, 18, 21, 25, 26, 28, 44]. However, a diminishing step size usually leads to slow convergence near the eventual limit and requires exhaustive experimentation to determine how rapidly the step size must decrease in order to prevent scenarios in which the step size becomes too small when the iterates are far from the eventual limit (e.g., determining the constants a and b in step sizes of the form $\mu(k) = a/(k + b)$).

A hybrid between the steepest descent method and the incremental gradient method was studied in [5]. The hybrid method starts as an incremental gradient method and gradually becomes the steepest descent. This method requires a tuning parameter, which controls the transition between the two methods, to gradually increase with k to ensure convergence. When the tuning parameter increases sufficiently fast with the number of iterations, it is shown that the rate of convergence is linear. However, the question of determining the rate of transition between the two methods still remains. For any fixed value of the tuning parameter, the hybrid method con-

verges to a limit cycle, unless a diminishing step size is used, similar to the standard incremental gradient method.

The choice of the aggregated gradient d^k (1.3) for generating an approximate descent direction was mentioned in [18] in the context of adaptive step size methods, which require repeated evaluations of either the complete objective function $f(x)$ or its gradient. This requirement renders the methods proposed in [18] inapplicable to problems in sensor networks of interest to us or any other applications which require decentralized implementation, as will be explained in section 3. In addition, as noted in [46], if $\nabla f_l(x)$, $l = 1, \ldots, L$, are not necessarily zero whenever $\nabla f(x) = 0$, the step size tends to zero, resulting in slow convergence.

The IAG method is closely related to Tseng's incremental gradient with momentum term [46], which is an incremental generalization of Polyak's heavy-ball method [34, p. 65] (also called the steepest descent with momentum term [7, p. 104]). Rewriting Tseng's method's update rule as

$$x^{k+1} = x^k - \mu(k) \sum_{l=0}^{k} \zeta^l \nabla f_{(k-l)_L}(x^{k-l}),$$

we see from (1.4) that the IAG method is a variation of this method with a truncated sum, $\zeta = 1$, and a constant step size. Similar to [18], the step size adaptation rule that leads to convergence in [46] requires repeated evaluations of the complete objective function $f(x)$ and its gradient. Hence, this method cannot be implemented in a distributed manner either. Furthermore, a linear convergence rate is established only under a certain growth property on the functions' gradients, which requires $\nabla f_l(x) = 0$, $l = 1, \ldots, L$, whenever $\nabla f(x) = 0$.

In contrast to the available methods, the IAG method has all four of the following properties: (a) it evaluates a single gradient per iteration, (b) it uses a constant step size, (c) it is convergent (Proposition 2.7), and (d) it has a global linear convergence rate for quadratic objective $f(x)$ (Proposition 2.8).

Finally, we note that the IAG method is reminiscent of other methods in various optimization problems, such as the incremental version of the Gauss–Newton method or the extended Kalman filter [2, 4, 15, 30], the distributed EM algorithm for maximum likelihood estimation [31, 33], the ordered subset and incremental optimization transfer for image reconstruction [1, 3, 10], and iterative methods for the convex feasibility problem [11, 12].

2. Convergence analysis. In this section we present convergence proofs for two different function classes: (I) restricted Lipschitz and (II) quadratic. Under a Lipschitz condition and a bounded gradient assumption on $f_l(x)$, $l = 1, \ldots, L$ (Assumptions 1 and 2), we obtain an upper bound on the limit inferior of $||\nabla f(x^k)||$, which depends linearly on the step size μ. By imposing additional restrictions on the function $f(x)$ (Assumptions 3 and 4), we prove pointwise convergence of the method. There are many functions that satisfy Assumptions 1–4. However, one important case does not satisfy these assumptions. This is the case when $f(x)$ and $f_l(x)$ are quadratic functions on $\mathbb{R}^p$. For this important case we provide a completely different convergence proof and show in addition that the convergence rate is globally linear.

For later reference, it will be useful to write (1.4) in a form known as the "gradient

method with errors" [8]:

$$x^{k+1} = x^k - \mu\frac{1}{L}\left[\sum_{l=0}^{L-1}\nabla f_{(k-l)_L}(x^k) + \sum_{l=0}^{L-1}\nabla f_{(k-l)_L}(x^{k-l}) - \sum_{l=0}^{L-1}\nabla f_{(k-l)_L}(x^k)\right]$$

$$= x^k - \mu\frac{1}{L}\left[\nabla f(x^k) + h^k\right], \tag{2.1}$$

where

$$h^k = \sum_{l=1}^{L-1}\left[\nabla f_{(k-l)_L}(x^{k-l}) - \nabla f_{(k-l)_L}(x^k)\right]$$

is the error term in the calculation of the gradient at x^k. Also note that for all $k \geq 2L$ and $1 \leq l \leq L$,

$$x^{k-l} - x^k = \mu\frac{1}{L}\left(d^{k-1} + d^{k-2} + \cdots + d^{k-l}\right).$$

2.1. Case I.

Assumption 1. $\nabla f_l(x)$, $l = 1, \ldots, L$, satisfy a Lipschitz condition in $\mathbb{R}^p$; i.e., there is a positive number M_1 such that for all $x, \overline{x} \in \mathbb{R}^p$, $||\nabla f_l(x) - \nabla f_l(\overline{x})|| \leq M_1||x - \overline{x}||$, $l = 1, \ldots, L$.

Assumption 1 implies that $\nabla f(x)$ also satisfies a Lipschitz condition; that is, for all $x, \overline{x} \in \mathbb{R}^p$, $||\nabla f(x) - \nabla f(\overline{x})|| \leq M_2||x - \overline{x}||$, where $M_2 = LM_1$.

Assumption 2. There exists a positive number M_3 such that for all $x \in \mathbb{R}^p$, $||\nabla f_l(x)|| \leq M_3$, $l = 1, \ldots, L$.

Assumption 2 implies that for all $x \in \mathbb{R}^p$, $||\nabla f(x)|| \leq M_4$, where $M_4 = LM_3$.

LEMMA 2.1. *Let $\{s_k\}_{k\geq 1}$ be a sequence of nonnegative real numbers satisfying for some fixed integer $L > 1$ and all $k \geq L$*

$$s_k \leq cQ(s_{k-1}, s_{k-2}, \ldots, s_{k-L+1}) + M,$$

where $0 < c < 1$, M is nonnegative, and $Q(s_{k-1}, s_{k-2}, \ldots, s_{k-L+1})$ is a linear form in the variables $s_{k-1}, s_{k-2}, \ldots, s_{k-L+1}$, whose coefficients are nonnegative and the sum of the coefficients equals one. Then $\limsup_{k\to\infty} s_k \leq \frac{M}{1-c}$.

Proof. Define the sequence $\{w_k\}_{k\geq 1}$ by $w_k = s_k$ for $1 \leq k \leq L-1$ and

$$w_k = cQ(w_{k-1}, w_{k-2}, \ldots, w_{k-L+1}) + M$$

for $k \geq L$. Since $s_k \leq w_k$ for all k, if $\lim_{k\to\infty} w_k = \frac{M}{1-c}$, then

$$\limsup_{k\to\infty} s_k \leq \limsup_{k\to\infty} w_k = \lim_{k\to\infty} w_k = \frac{M}{1-c}.$$

To show that $\lim_{k\to\infty} w_k = \frac{M}{1-c}$, define the sequence $\{v_k\}_{k\geq 1}$ by $v_k = s_k - \frac{M}{1-c}$ for $1 \leq k \leq L-1$ and

$$v_k = cQ(v_{k-1}, v_{k-2}, \ldots, v_{k-L+1})$$

for $k \geq L$. By this construction,

$$w_L = cQ\left(\frac{M}{1-c} + v_{L-1}, \frac{M}{1-c} + v_{L-2}, \ldots, \frac{M}{1-c} + v_1\right) + M$$

$$= c\frac{M}{1-c} + cQ(v_{L-1}, v_{L-2}, \ldots, v_1) + M = \frac{M}{1-c} + v_L,$$

and, by induction, $w_k = \frac{M}{1-c} + v_k$ for all $k > L$. Therefore, if $\lim_{k\to\infty} v_k = 0$, then $\lim_{k\to\infty} w_k = \frac{M}{1-c}$. To show that $\lim_{k\to\infty} v_k = 0$, set $A = \max\{|v_1|, |v_2|, \ldots, |v_{L-1}|\}$. Hence,

$$|v_L| = c|Q(v_{L-1}, v_{L-2}, \ldots, v_1)| \leq cQ(|v_{L-1}|, |v_{L-2}|, \ldots, |v_1|) \leq cA.$$

Similarly, $|v_{L+1}| \leq cA$, and in general $|v_k| \leq cA$ for all $k \geq L$. Consider now v_{2L}. Since $\max\{|v_{2L-1}|, |v_{2L-2}|, \ldots, |v_{L+1}|\} \leq cA$, we have

$$|v_{2L}| = c|Q(v_{2L-1}, v_{2L-2}, \ldots, v_{L+1})| \leq cQ(|v_{2L-1}|, |v_{2L-2}|, \ldots, |v_{L+1}|) \leq c^2 A,$$

and in general $|v_k| \leq c^2 A$ for all $k \geq 2L$. Similarly, we obtain $|v_k| \leq c^n L$ for all $k \geq nL$. Since $0 < c < 1$, we have $\lim_{n\to\infty} c^n = 0$, and therefore $\lim_{k\to\infty} v_k = 0$. □

Remark 1. Lemma 2.1 can also be proven using concepts from dynamical systems. The sequence w_k is the output of an autoregressive linear system

$$w_k = c\sum_{l=1}^{L-1} \alpha_k w_{k-l} + Mu(k-L),$$

where $u(k)$ is the unit step function which equals one when $k \geq 0$ and zero otherwise, with initial condition $w_k = s_k$ for $1 \leq k \leq L-1$. Since the coefficients of the linear form are all positive and sum to one, and $0 < c < 1$, it is possible to show that the system is stable (bounded input bounded output) and the steady state response is $\frac{M}{1-c}$ [35], i.e., $\lim_{k\to\infty} w_k = \frac{M}{1-c}$.

LEMMA 2.2. *Under Assumption* 1, *if* $||\nabla f(x^k)|| > \frac{||h^k||}{1-2\mu M_1}$, *and* $0 < 1-2\mu M_1 < 1$, *then* $f(x^k) > f(x^{k+1})$.

Proof. Assume that $||\nabla f(x^k)|| > \frac{||h^k||}{1-2\mu M_1}$. Then

$$\begin{aligned} ||d^k||^2 = ||\nabla f(x^k) + h^k||^2 &\leq 2||\nabla f(x^k)||^2 + 2||h^k||^2 \\ &< 2||\nabla f(x^k)||^2 + 2\frac{||h^k||^2}{1-2\mu M_1} < 4||\nabla f(x^k)||^2. \end{aligned}$$

By [6, Prop. A.24], if Assumption 1 holds, then

$$f(x+y) - f(x) \leq y'\nabla f(x) + \frac{1}{2}M_2||y||^2.$$

Hence

$$\begin{aligned} f(x^k) - f(x^{k+1}) &= f(x^k) - f\left(x^k - \mu\frac{1}{L}d^k\right) \\ &\geq \mu\frac{1}{L}d^{k\prime}\nabla f(x^k) - \frac{1}{2}M_2\mu^2\frac{1}{L^2}||d^k||^2 \\ &> \mu\frac{1}{L}(\nabla f(x^k) + h^k)'\nabla f(x^k) - \frac{1}{2}M_2\mu^2\frac{1}{L^2}4||\nabla f(x^k)||^2 \\ &= \mu\frac{1}{L}||\nabla f(x^k)||^2 + \mu\frac{1}{L}h^{k\prime}\nabla f(x^k) - 2M_2\mu^2\frac{1}{L^2}||\nabla f(x^k)||^2 \\ &\geq \mu\frac{1}{L}||\nabla f(x^k)||^2 - \mu\frac{1}{L}||h^k|| \cdot ||\nabla f(x^k)|| - 2M_2\mu^2\frac{1}{L^2}||\nabla f(x^k)||^2 \\ &= \frac{\mu}{L}||\nabla f(x^k)||(1-2\mu M_1)\left(||\nabla f(x^k)|| - \frac{||h^k||}{1-2\mu M_1}\right) \\ &> 0. \end{aligned}$$

□

LEMMA 2.3. *Set $\delta_0 = \mu M_2 M_3$. Under Assumptions 1 and 2, if $\mu M_2 < 1$, there exists K such that for all $k > K$, $||h^k|| < \delta_0$.*

Proof.

$$\begin{aligned}
||h^k|| &\leq \sum_{l=1}^{L-1} ||\nabla f_{(k-l)_L}(x^{k-l}) - \nabla f_{(k-l)_L}(x^k)|| \\
&\leq M_1 \sum_{l=1}^{L-1} ||x^{k-l} - x^k|| \\
&= \mu M_1 \frac{1}{L} \sum_{l=1}^{L-1} ||d^{k-1} + d^{k-2} + \cdots + d^{k-l}|| \\
&\leq \mu M_1 \frac{1}{L} \sum_{l=1}^{L-1} \left(||d^{k-1}|| + ||d^{k-2}|| + \cdots + ||d^{k-l}|| \right) \\
&= \mu M_1 \frac{1}{L} \left[(L-1)||d^{k-1}|| + (L-2)||d^{k-2}|| + \cdots + ||d^{k-L+1}|| \right] \\
&= \mu M_1 \frac{1}{L} \frac{L(L-1)}{2} \left[\frac{(L-1)||d^{k-1}|| + (L-2)||d^{k-2}|| + \cdots + ||d^{k-L+1}||}{L(L-1)/2} \right] \\
&= \mu M_1 \frac{L-1}{2} Q(||d^{k-1}||, ||d^{k-2}||, \ldots, ||d^{k-L+1}||),
\end{aligned}$$

where $Q(||d^{k-1}||, ||d^{k-2}||, \ldots, ||d^{k-L+1}||)$ is a linear form in the variables $||d^{k-1}||$, $||d^{k-2}||, \ldots, ||d^{k-L+1}||$ whose coefficients, $\frac{L-1}{L(L-1)/2}, \frac{L-2}{L(L-1)/2}, \ldots, \frac{1}{L(L-1)/2}$, sum to one. Next, we use $||d^k|| = ||\nabla f(x^k) + h^k|| \leq ||\nabla f(x^k)|| + ||h^k||$ to obtain

$$\begin{aligned}
||h^k|| &\leq \mu M_1 \frac{L-1}{2} Q(||h^{k-1}||, ||h^{k-2}||, \ldots, ||h^{k-L+1}||) \\
&\quad + \mu M_1 \frac{L-1}{2} Q(||\nabla f(x^{k-1})||, ||\nabla f(x^{k-2})||, \ldots, ||\nabla f(x^{k-L+1})||) \\
&\leq \mu M_1 \frac{L-1}{2} Q(||h^{k-1}||, ||h^{k-2}||, \ldots, ||h^{k-L+1}||) + \mu M_1 \frac{L-1}{2} M_3 \\
&< \mu \frac{M_2}{2} Q(||h^{k-1}||, ||h^{k-2}||, \ldots, ||h^{k-L+1}||) + \mu \frac{M_2}{2} M_3,
\end{aligned}$$

where Assumption 2 was used in the second to last inequality. Hence, by Lemma 2.1, since $0 < \mu \frac{M_2}{2} < 1/2$, $\limsup_{k\to\infty} ||h^k|| \leq \frac{\mu \frac{M_2}{2} M_3}{1 - \mu \frac{M_2}{2}}$. By using $\mu \frac{M_2}{2} < 1/2$, we obtain $\limsup_{k\to\infty} ||h^k|| < \mu M_2 M_3$ and the lemma follows. ∎

PROPOSITION 2.4. *Under Assumptions 1 and 2, if $f(x)$ is bounded from below and $\mu \max\{2M_1, M_2\} < 1$, then*

$$\liminf_{k\to\infty} ||\nabla f(x^k)|| \leq \frac{2M_2 M_3}{1 - 2\mu M_1} \mu.$$

Proof. The proof is similar to the proof of Theorem 2.1 in [44]. ∎

Next, by imposing two additional assumptions, we prove that the IAG method converges with a constant step size to the minimum point of $f(x)$.

Assumption 3. $f(x)$ has a unique global minimum at x^*. The Hessian $\nabla^2 f(x)$ is continuous and positive definite at x^*.

Assumption 4. For any sequence $\{t^k\}_{k=1}^{\infty}$ in $\mathbb{R}^p$, if $\lim_{k\to\infty} f(t^k) = f(x^*)$ or $\lim_{k\to\infty} ||\nabla f(t^k)|| = 0$, then $\lim_{k\to\infty} t^k = x^*$.

There is an equivalent form of Assumption 4: For each neighborhood $\mathcal{U}$ of x^* there exists $\eta > 0$ such that if $f(x) - f(x^*) < \eta$ or $||\nabla f(x)|| < \eta$, then $x \in \mathcal{U}$.

Remark 2. Assumptions 3 and 4 are stronger than the assumptions usually made on $f(x)$ in the literature (see [8] for a summary of the available convergence proofs and the assumptions they require). However, our results hold for a constant step size and do not require that $\nabla f_l(x) = 0$, $l = 1, \ldots, L$, whenever $\nabla f(x) = 0$. In addition, note that there are nonconvex functions that satisfy Assumption 4. However, if $f(x)$ is strictly convex and takes a minimum in the interior of its domain ($\mathbb{R}^p$), then Assumption 4 is automatically satisfied. In particular, if $f(x)$ satisfies Assumption 3 and is strictly convex, then Assumption 4 is satisfied. In fact, the implication $\lim_{k\to\infty} f(t^k) = f(x^*) \Rightarrow \lim_{k\to\infty} t^k = x^*$ is the statement of Corollary 27.2.2 from [41]. The implication $\lim_{k\to\infty} ||\nabla f(t^k)|| = 0 \Rightarrow \lim_{k\to\infty} t^k = x^*$ can be obtained as follows: Consider the function $\nabla f : \mathbb{R}^p \to \mathbb{R}^p$. The derivative $(\nabla f)'$ of this function is the Hessian $\nabla^2 f$. Since $f(x)$ satisfies Assumption 3 and is strictly convex, $\det(\nabla f)' \neq 0$. Therefore, by the inverse function theorem, there are open neighborhoods V of $x^* \in \mathbb{R}^p$ and W of $0 \in \mathbb{R}^p$ such that $\nabla f : V \to W$ has a continuous inverse $\gamma : W \to V$. Let $\{t^k\}_{k=1}^{\infty}$ be a sequence such that $\lim_{k\to\infty} ||\nabla f(t^k)|| = 0$. Then there exists k_0 such that $\nabla f(t^k) \in W$ for all $k \geq k_0$. By Theorem B on page 99 in [40], since $f(x)$ is strictly convex, ∇f is one-to-one; i.e., if $x \neq y$, then $\nabla f(x) \neq \nabla f(y)$. It follows that $t^k \in V$ for all $k \geq k_0$. Now we have

$$\begin{aligned}\lim_{k\to\infty} t^k &= \lim_{k\to\infty} \gamma\left(\nabla f(t^k)\right) \\ &= \gamma\left(\lim_{k\to\infty} \nabla f(t^k)\right) \\ &= \gamma(0) = x^*.\end{aligned}$$

Remark 3. Unimodal functions which are convex in the neighborhood of their minimum and have bounded gradient are common in robust estimation [20]. An example of a robust estimation objective function that satisfies Assumptions 1–4 is given in section 4.1. Another important function which satisfies Assumptions 1–4 is the objective function minimized by the LogitBoost algorithm [16] (or adaptive logistic regression). To explain the components which are used to construct this objective function we include a short description (taken from [14]) of the supervised learning problem, and in particular, the problem of combining weak features. Let $\{z_l, y_l\}_{l=1}^{L}$ be a set of training examples, where each instance z_l takes values in an instance domain $\mathcal{Z}$, and each y_l, called the label, takes values in $\{-1, +1\}$. Given a set of p real-valued functions on $\mathcal{Z}$, $h_1, h_2, \ldots, h_p$ called features, the goal is to find a vector $x \in \mathbb{R}^p$ for which the sign of $g_x(z_l) = \sum_{i=1}^{p} x_i h_i(z_l)$ is a good predictor of y_l for $l = 1, \ldots, L$. Let M be the $L \times p$ matrix whose (l, i) element is $h_i(z_l)$. The objective function $f(x) : \mathbb{R}^p \to \mathbb{R}$ minimized by the LogitBoost algorithm [14] is given by

$$f(x) = \sum_{l=1}^{L} \log\left[1 + \exp\left(-y_l [Mx]_l\right)\right], \tag{2.2}$$

where $[Mx]_l$ is the lth element of the vector Mx. It can be motivated as being a convex surrogate to the nonconvex and nondifferentiable $0-1$ loss function

$$f(x) = \sum_{l=1}^{L} I(g_x(z_l)y_l \leq 0),$$

which is the number of labels that are not predicted correctly by the sign of $g_x(z_l)$, or through the maximum likelihood method for estimating the conditional probability of y_l given z_l. It is shown below that in the nonseparable case, i.e., when there exists no value of x for which $\operatorname{sign}(g_x(z_l)) = y_l$, for $l = 1, \ldots, L$, and when the features are linearly independent on the training set, i.e., $\operatorname{rank} M = p$, the function $f(x)$ (2.2) satisfies Assumptions 1–4:

$$\frac{\partial}{\partial x_j} \log\left[1 + \exp\left(-y_l[Mx]_l\right)\right] = \frac{\exp\left(-y_l[Mx]_l\right)}{1 + \exp\left(-y_l[Mx]_l\right)} \left(-y_l h_j(z_l)\right) \leq |h_j(z_l)|.$$

Hence Assumption 2 holds:

$$\begin{aligned}\frac{\partial^2}{\partial x_j \partial x_k} \log\left[1 + \exp\left(-y_l[Mx]_l\right)\right] &= \frac{\exp\left(-y_l[Mx]_l\right)}{\left[1 + \exp\left(-y_l[Mx]_l\right)\right]^2} h_j(z_l)h_k(z_l) \\ &\leq |h_j(z_l)h_k(z_l)|.\end{aligned}$$

Hence Assumption 1 holds. Let $d_l(x) = \exp\left(-y_l[Mx]_l\right) / \left[1 + \exp\left(-y_l[Mx]_l\right)\right]^2 > 0$. Then

$$\frac{\partial^2 f(x)}{\partial x_j \partial x_k} = \sum_{l=1}^{L} d_l(x) M_{lj} M_{lk}.$$

To show that $\nabla f(x)$ is positive definite for all x, consider $\zeta^T \nabla f(x) \zeta$ for some vector $\zeta \in \mathbb{R}^p$:

$$\zeta^T \nabla f(x) \zeta = \sum_{j,k=1}^{p} \sum_{l=1}^{L} d_l(x) M_{lk} M_{lj} \zeta_k \zeta_j = \sum_{l=1}^{L} d_l(x) \left([M\zeta]_l\right)^2 \geq 0$$

with equality if and only if $\zeta = 0$, by the assumption that $\operatorname{rank} M = p$. Hence the function $f(x)$ is strictly convex. Assume the training set $\{z_l, y_l\}_{l=1}^{L}$ is nonseparable with respect to the features $h_1, h_2, \ldots, h_p$; i.e., for every x there exists at least one l for which $y_l[Mx]_l < 0$. For any given $x \neq 0$ let $I_1(x) = \{l : y_l[Mx]_l < 0\}$, $I_2(x) = \{l : y_l[Mx]_l = 0\}$, and $I_3(x) = \{l : y_l[Mx]_l > 0\}$, and note that $I_1(x)$ is nonempty by assumption. For a positive scalar c, we can write $f(cx)$ as the sum of three summations:

$$\begin{aligned}f(cx) = &\sum_{l \in I_1(x)} \log\left\{1 + \exp\left[-cy_l \sum_{i=1}^{p} x_i h_i(z_l)\right]\right\} \\ &+ \sum_{l \in I_2(x)} \log 2 \\ &+ \sum_{l \in I_3(x)} \log\left\{1 + \exp\left[-cy_l \sum_{i=1}^{p} x_i h_i(z_l)\right]\right\}.\end{aligned}$$

When $c \to \infty$,

$$\sum_{l \in I_1(x)} \log \left\{ 1 + \exp \left[-cy_l \sum_{i=1}^{p} x_i h_i(z_l) \right] \right\} \to \infty$$

and

$$\sum_{l \in I_3(x)} \log \left\{ 1 + \exp \left[-cy_l \sum_{i=1}^{p} x_i h_i(z_l) \right] \right\} \to 0.$$

Therefore, $\lim_{c \to \infty} f(cx) = \infty$ for all $x \neq 0$. This implies that $f(x)$ has no directions of recession. A direction of recession is a nonzero vector x^1 such that $f(x^2 + cx^1)$ is a nonincreasing function of the scalar c for every choice of vector x^2. Hence by Theorem 27.1(d) in [41, p. 265] the minimum set of $f(x)$ is nonempty. The minimum is unique by the strict convexity of $f(x)$. Therefore, Assumption 3 is also satisfied, and the strict convexity, together with Assumption 3, implies Assumption 4 as well.

The following lemma is well known.

LEMMA 2.5. *Under Assumption* 3, *there exists a neighborhood* $\mathcal{U}$ *of* x^* *and positive constants* A_1, A_2, B_1, B_2 *such that for all* $x \in \mathcal{U}$,

$$A_1 ||x - x^*||^2 \leq f(x) - f(x^*) \leq B_1 ||x - x^*||^2, \tag{2.3}$$

$$A_2 ||x - x^*||^2 \leq ||\nabla f(x)||^2 \leq B_2 ||x - x^*||^2. \tag{2.4}$$

Let $\mathcal{U}$ be a neighborhood of x^* for which inequalities (2.3) and (2.4) hold. By Assumption 4 there exists $\eta > 0$ such that $x \in \mathcal{U}$ if $f(x) - f(x^*) < \eta$ or $||\nabla f(x)|| < \eta$.

LEMMA 2.6. *Set* $M_5 = \max\{3\sqrt{\frac{B_1 B_2}{A_1 A_2}}, \frac{2}{1-2\mu M_1}\}$ *and* $\lambda = \mu M_2 M_5$. *Under Assumptions* 1, 3, *and* 4, *if there exist positive numbers* n_1 *and* δ *such that* $||h^k|| < \delta$ *for every* $k \geq n_1$, $3\delta < \eta$, $\frac{9B_1}{A_2}\delta^2 < \eta$, *and* $9\mu M_1 < 1$, *then*

(i) *there exists a number* k_1 *such that* $||\nabla f(x^k)|| < M_5\delta$ *and* $||d^k|| < 2M_5\delta$ *for every* $k \geq k_1$, *and*

(ii) *there exists a number* n_2 *such that* $||h^k|| < \lambda\delta$ *for every* $k \geq n_2$.

Proof. First, we show that there exists k such that $k \geq n_1$ and $||\nabla f(x^k)|| < \frac{2\delta}{1-2\mu M_1}$. In fact, if $||\nabla f(x^k)|| \geq \frac{2\delta}{1-2\mu M_1}$ for all $k \geq n_1$, then $||\nabla f(x^k)|| > \frac{2||h^k||}{1-2\mu M_1} \geq \frac{||h^k||}{1-2\mu M_1}$ for all $k \geq n_1$. By Lemma 2.2, the sequence $\{f(x^k)\}_{k=n_1}^{\infty}$is decreasing. Since it is bounded from below by $f(x^*)$, there exists $\lim_{k \to \infty} f(x^k)$. By replacing δ_0 with δ and $\max\{K_1, K_2\}$ with n_1 at the last argument of the proof of Proposition 2.4, we obtain a contradiction.

Let k_1 be the smallest natural number such that $k_1 \geq n_1$ and $||\nabla f(x^{k_1})|| \leq \frac{2\delta}{1-2\mu M_1}$. Without loss of generality, assume there exists k_2, the smallest natural number such that $k_2 > k_1$ and $||\nabla f(x^{k_2})|| > \frac{2\delta}{1-2\mu M_1}$. Let k_3 be the smallest natural number such that $k_3 > k_2$ and $||\nabla f(x^{k_3})|| \leq \frac{2\delta}{1-2\mu M_1}$. Let k_4 be the smallest natural number such that $k_4 > k_3$ and $||\nabla f(x^{k_4})|| > \frac{2\delta}{1-2\mu M_1}$. We define $k_5, k_6, \ldots$ in a similar manner.

For every natural m,

$$||d^{k_{2m}-1}|| \leq ||\nabla f(x^{k_{2m}-1})|| + ||h^{k_{2m}-1}|| \leq \frac{2\delta}{1-2\mu M_1} + \delta \leq \frac{3\delta}{1-2\mu M_1},$$

$$||x^{k_{2m}} - x^{k_{2m}-1}|| = \mu \frac{1}{L} ||d^{k_{2m}-1}|| \leq \frac{3\mu/L}{1-2\mu M_1}\delta,$$

and

$$\begin{aligned}||\nabla f(x^{k_{2m}})|| &\leq ||\nabla f(x^{k_{2m}}) - \nabla f(x^{k_{2m}-1})|| + ||\nabla f(x^{k_{2m}-1})|| \\ &\leq M_2||x^{k_{2m}} - x^{k_{2m}-1}|| + \frac{2\delta}{1-2\mu M_1} \\ &\leq M_2\frac{3\mu/L}{1-2\mu M_1}\delta + \frac{2}{1-2\mu M_1}\delta \\ &= \frac{2+3\mu M_1}{1-2\mu M_1}\delta < 3\delta,\end{aligned}$$

where we used $\mu < \frac{1}{9M_1}$ to obtain the last inequality.

Since $||\nabla f(x^{k_{2m}})|| < 3\delta < \eta$, $x^{k_{2m}} \in \mathcal{U}$, and we can use Lemma 2.5. We obtain

$$f(x^{k_{2m}}) - f(x^*) \leq B_1||x^{k_{2m}} - x^*|| \leq \frac{B_1}{A_2}||\nabla f(x^{k_{2m}})||^2 < \frac{B_1}{A_2}9\delta^2.$$

Let k be such that $k_{2m} \leq k < k_{2m+1}$. Then, by Lemma 2.2,

$$f(x^k) - f(x^*) < f(x^{k_{2m}}) - f(x^*) < 9\frac{B_1}{A_2}\delta^2.$$

Since $f(x^k) - f(x^*) < 9\frac{B_1}{A_2}\delta^2 < \eta$, $x^k \in \mathcal{U}$, and we can use Lemma 2.5. We obtain

$$||\nabla f(x^k)||^2 \leq B_2||x^k - x^*||^2 \leq \frac{B_2}{A_1}\left[f(x^k) - f(x^*)\right] < 9\frac{B_1B_2}{A_1A_2}\delta^2.$$

Thus, if k satisfies $k_{2m} \leq k < k_{2m+1}$, we have $||\nabla f(x^k)|| < 3\sqrt{\frac{B_1B_2}{A_1A_2}}\delta$. If k satisfies $k_{2m-1} \leq k < k_{2m}$, we have $||\nabla f(x^k)|| < \frac{2}{1-2\mu M_1}\delta$. Therefore for each $k \geq k_1$, $||\nabla f(x^k)|| < M_5\delta$, and therefore

$$||d^k|| \leq ||\nabla f(x^k)|| + ||h^k|| \leq M_5\delta + \delta < 2M_5\delta.$$

Thus, if $k \geq k_1$, we have

$$\begin{aligned}||\nabla f(x^k)|| &< M_5\delta, \\ ||d^k|| &< 2M_5\delta.\end{aligned} \tag{2.5}$$

This proves the first part of the lemma.

To prove the second part, we take $n_2 = k_1 + L - 1$. If $k \geq n_2$, then not only x^k but also $L-1$ previous terms of the sequence $\{x^k\}$ satisfy inequalities (2.5). Therefore, by following the steps in the proof of Proposition 2.4, we have for $k \geq n_2$

$$\begin{aligned}||h^k|| &\leq \mu M_1\frac{1}{L}\sum_{l=1}^{L-1}\left(||d^{k-1}|| + ||d^{k-2}|| + \cdots + ||d^{k-l}||\right) \\ &< \mu M_1\frac{1}{L}2M_5\delta\sum_{l=1}^{L-1}\sum_{m=1}^{l}1 = \mu M_1\frac{1}{L}2M_5\delta\frac{L(L-1)}{2} \\ &< \mu M_2M_5\delta = \lambda\delta.\end{aligned}$$

Thus $||h^k|| < \lambda\delta$. This proves the second part of Lemma 2.6. ∎

Remark 4. A direct result of Lemma 2.6 is that under Assumptions 1–4, $||h^k|| \to 0$ is a sufficient condition for the convergence of x^k, generated by any gradient method with errors (2.1), to x^*.

PROPOSITION 2.7. *Under Assumptions* 1, 2, 3, *and* 4, *if* $\mu < \min\{\frac{1}{9M_1}, \frac{1}{M_2M_5}, \frac{\eta}{3M_1M_3}, \frac{1}{3M_2M_3}\sqrt{\frac{A_2\eta}{B_1}}\}$, *then* $\lim_{k\to\infty} x^k = x^*$.

Proof. We prove Proposition 2.7 by repeated use of Lemma 2.6. We start with $\delta = \delta_0$. By applying Lemma 2.3, there exists K such that for all $k > K$, $||h^k|| < \delta_0$. After applying Lemma 2.6 r times we get a number n_r such that $||h^k|| < \delta_0\lambda^r$, $||\nabla f(x^k)|| < M_5\delta_0\lambda^r$, and $||d^k|| < 2M_5\delta_0\lambda^r$ for $k \geq n_r$. The inequality $\mu < \frac{1}{M_2M_5}$ is equivalent to $0 < \lambda < 1$. Hence, $\lim_{k\to\infty} ||h^k|| = 0$, $\lim_{k\to\infty} ||d^k|| = 0$, and $\lim_{k\to\infty} ||\nabla f(x^k)|| = 0$, and by Assumption 4, $\lim_{k\to\infty} x^k = x^*$.

Note that the inequality $\mu < \frac{1}{9M_1}$ was used in the proof of Lemma 2.6, and the inequalities $\mu < \frac{\eta}{3M_2M_3}$ and $\mu < \frac{1}{3M_2M_3}\sqrt{\frac{A_2\eta}{B_1}}$ are equivalent to $3\delta_0 < \eta$ and $\frac{9B_1}{A_2}\delta_0^2 < \eta$, respectively. ∎

2.2. Case II: Quadratic case. In [25] it is shown that when applied to the objective function

$$f(x) = \frac{1}{2}(x - c_1)^2 + \frac{1}{2}(x - c_2)^2,$$

the standard incremental gradient method with a constant step size

$$x^{k+1} = x^k - \mu\nabla f_{(k)_L}(x^k)$$

converges to a limit cycle with limit points

$$x_1^*(\mu) = \frac{(1-\mu)c_1 + c_2}{2-\mu}, \qquad x_2^*(\mu) = \frac{(1-\mu)c_2 + c_1}{2-\mu}$$

whenever $0 < \mu < 1$. When implementing the IAG method one obtains

$$\begin{aligned} x^{k+1} &= x^k - \frac{\mu}{2}\left[\left(x^k - c_{(k)_2}\right) + \left(x^{k-1} - c_{(k-1)_2}\right)\right] \\ &= x^k - \frac{\mu}{2}\left[x^k + x^{k-1} - (c_1 + c_2)\right]. \end{aligned}$$

Subtracting $x^* = (c_1 + c_2)/2$, the unique minimum of $f(x)$, from both sides and denoting the error at the kth iteration by $e^k = x^k - x^*$ lead to the following error form:

$$e^{k+1} = e^k - \frac{\mu}{2}\left[e^k + e^{k-1}\right].$$

The characteristic polynomial of this linear system is $\lambda^2 - (1 - \mu/2)\lambda + \mu/2$, and it is easy to show that the roots of this polynomial are inside the unit circle whenever $0 < \mu < 2$. Hence, when $0 < \mu < 2$, $e^k \to 0$; i.e., x^k converges to the unique minimum, in contrast to the standard incremental gradient method.

More generally, suppose that the functions f_l, $l = 1, \ldots, L$, have the following form:

$$f_l(x) = \frac{1}{2}x'Q_lx - c_l'x, \quad l = 1, \ldots, L, \tag{2.6}$$

where Q_l are given symmetric matrices, c_l are given vectors, and $\sum_{l=1}^{L} Q_l$ is positive definite. Under this assumption, the function $f(x) = \sum_{l=1}^{L} f_l(x)$ is strictly convex, having its minimum point at

$$x^* = \left(\sum_{l=1}^{L} Q_l\right)^{-1} \sum_{l=1}^{L} c_l, \tag{2.7}$$

and x^* is the only stationary point of $f(x)$.

PROPOSITION 2.8. *For sufficiently small μ, $\lim_{k\to\infty} x^k = x^*$, and the rate of convergence of the IAG method* (1.4) *is linear.*

Proof. Plugging (2.6) into (1.4), the IAG method becomes

$$x^{k+1} = x^k - \mu\left[\sum_{l=0}^{L-1} Q_{(k-l)_L} x^{k-l} - c_{(k-l)_L}\right] = x^k - \mu \sum_{l=0}^{L-1} Q_{(k-l)_L} x^{k-l} + \mu c,$$

where $c = \sum_{l=1}^{L} c_l$, and the factor $\frac{1}{L}$ was absorbed into μ to simplify the notation. Subtracting x^* (2.7) from both sides and adding and subtracting x^* inside the parentheses, we obtain

$$x^{k+1} - x^* = x^k - x^* - \mu \sum_{l=0}^{L-1} Q_{(k-l)_L}(x^{k-l} - x^* + x^*) + \mu c.$$

Denoting the error at the kth iteration by $e^k = x^k - x^*$ and the substitution of (2.7) for x^* lead to the following error form:

$$e^{k+1} = e^k - \mu \sum_{l=0}^{L-1} Q_{(k-l)_L} e^{k-l}.$$

This relation between a new error and the previous errors can be seen as a periodically time varying linear system. To analyze its stability, which will lead to the convergence result, it is useful to consider L iterations as one iteration [29]. This can be seen as downsampling the original system by a factor of L, which leads to a time invariant system of a lower sampling rate. Without loss of generality, consider the case where $k = NL$ for some integer N; i.e., $k+1$ corresponds to the first iteration of a new cycle. In this case we have

$$\begin{aligned} e^{k+1} &= e^k - \mu \sum_{l=0}^{L-1} Q_{(k-l)_L} e^{k-l} = e^k - \mu \begin{bmatrix} Q_L & Q_{L-1} & Q_{L-2} & \dots & Q_1 \end{bmatrix} \overline{e}^k \\ &= \begin{bmatrix} I_p - \mu Q_L & -\mu Q_{L-1} & -\mu Q_{L-2} & \dots & -\mu Q_1 \end{bmatrix} \overline{e}^k, \end{aligned}$$

where I_p is the $p \times p$ identity matrix and

$$\overline{e}^k = \begin{bmatrix} e^k \\ e^{k-1} \\ \vdots \\ e^{k-L+1} \end{bmatrix}.$$

Similarly,

$$\begin{aligned} e^{k+2} &= e^{k+1} - \mu \sum_{l=0}^{L-1} Q_{(k+1-l)L} e^{k+1-l} \\ &= e^{k+1} - \mu \begin{bmatrix} Q_1 & Q_L & Q_{L-1} & \dots & Q_2 \end{bmatrix} \overline{e}^{k+1} \\ &= \begin{bmatrix} I_p - \mu Q_1 & -\mu Q_L & -\mu Q_{L-1} & \dots & -\mu Q_2 \end{bmatrix} \overline{e}^{k+1}, \end{aligned}$$

and finally

$$\begin{aligned} e^{k+L} &= e^{k+L-1} - \mu \sum_{l=0}^{L-1} Q_{(k+L-1-l)L} e^{k+L-1-l} \\ &= e^{k+L-1} - \mu \begin{bmatrix} Q_{L-1} & Q_{L-2} & Q_{L-3} & \dots & Q_L \end{bmatrix} \overline{e}^{k+L-1} \\ &= \begin{bmatrix} I_p - \mu Q_{L-1} & -\mu Q_{L-2} & -\mu Q_{L-3} & \dots & -\mu Q_L \end{bmatrix} \overline{e}^{k+L-1}. \end{aligned}$$

This leads to the relation

$$\overline{e}^{k+L} = M_L \overline{e}^{k+L-1},$$

where

$$M_L = \begin{bmatrix} I_p - \mu Q_{L-1} & -\mu Q_{L-2} & \dots & -\mu Q_1 & -\mu Q_L \\ I_p & 0_p & \dots & 0_p & 0_p \\ 0_p & I_p & \dots & 0_p & 0_p \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0_p & 0_p & \dots & I_p & 0_p \end{bmatrix},$$

where 0_p denotes the $p \times p$ zero matrix. Taking another step we have

$$\overline{e}^{k+L} = M_L M_{L-1} \overline{e}^{k+L-2},$$

where

$$M_{L-1} = \begin{bmatrix} I_p - \mu Q_{L-2} & -\mu Q_{L-3} & \dots & -\mu Q_L & -\mu Q_{L-1} \\ I_p & 0_p & \dots & 0_p & 0_p \\ 0_p & I_p & \dots & 0_p & 0_p \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0_p & 0_p & \dots & I_p & 0_p \end{bmatrix},$$

and finally, by induction,

$$\overline{e}^{k+L} = M_L M_{L-1} \dots M_1 \overline{e}^{k},$$

where

$$M_1 = \begin{bmatrix} I_p - \mu Q_L & -\mu Q_{L-1} & \dots & -\mu Q_2 & -\mu Q_1 \\ I_p & 0_p & \dots & 0_p & 0_p \\ 0_p & I_p & \dots & 0_p & 0_p \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0_p & 0_p & \dots & I_p & 0_p \end{bmatrix}.$$

Denoting $M = M_L M_{L-1} \dots M_1$, we have $\overline{e}^{k+L} = M\overline{e}^k$, and in general $\overline{e}^{k+nL} = M^n\overline{e}^k$. Therefore, if for sufficiently small $\mu > 0$ the eigenvalues of M are inside the unit circle, then $\lim_{n\to\infty} \overline{e}^{k+nL} = 0_{pL\times 1}$, where $0_{pL\times 1}$ is a $pL \times 1$ zero vector; i.e., the method converges to the minimum of the function $f(x)$ and the convergence rate is linear.

To prove that the eigenvalues of M are inside the unit circle, set

$$A = \begin{bmatrix} I_p & 0_p & \dots & 0_p & 0_p \\ I_p & 0_p & \dots & 0_p & 0_p \\ 0_p & I_p & \dots & 0_p & 0_p \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0_p & 0_p & \dots & I_p & 0_p \end{bmatrix}$$

and

$$B_k = \begin{bmatrix} Q_{(k-1)_L} & Q_{(k-2)_L} & \dots & Q_{(k+1)_L} & Q_k \\ 0_p & 0_p & \dots & 0_p & 0_p \\ 0_p & 0_p & \dots & 0_p & 0_p \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0_p & 0_p & \dots & 0_p & 0_p \end{bmatrix}, \quad k = 1, \dots, L,$$

so that $M_k = A - \mu B_k$ and $M = (A - \mu B_L)(A - \mu B_{L-1}) \dots (A - \mu B_1)$. Hence,

$$M = A^L - \mu\big(B_L A^{L-1} + A B_{L-1} A^{L-2} + A^2 B_{L-2} A^{L-3} + \cdots + A^{L-2} B_2 A + A^{L-1} B_1\big) + \mu^2 C(\mu),$$

where $C(\mu)$ is an $Lp \times Lp$ matrix whose elements are polynomials in μ.

Note that premultiplying a matrix by A will duplicate the first row of $p\times p$ matrices and will shift the rest of the rows down, discarding the last p rows. Postmultiplying by A will add the second column of $p \times p$ matrices to the first one and will shift the rest of the columns to the left, inserting a block of $p \times p$ zero matrices to the last column. It follows that

$$A^L = \begin{bmatrix} I_p & 0_p & \dots & 0_p & 0_p \\ I_p & 0_p & \dots & 0_p & 0_p \\ I_p & 0_p & \dots & 0_p & 0_p \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ I_p & 0_p & \dots & 0_p & 0_p \end{bmatrix}$$

and

$$A^{L-k} B^k A^{k-1} = \begin{bmatrix} W_1(k) & 0_{(L-k+1)p\times(k-1)p} \\ 0_{(k-1)p\times(L-k+1)p} & 0_{(k-1)p\times(k-1)p} \end{bmatrix},$$

where $W_1(k)$ is a $(L-k+1)p \times (L-k+1)p$ matrix whose elements are

$$W_1(k) = \begin{bmatrix} \sum_{l=0}^{k-1} Q_{(l)_L} & Q_{L-1} & \dots & Q_k \\ \vdots & \vdots & & \vdots \\ \sum_{l=0}^{k-1} Q_{(l)_L} & Q_{L-1} & \dots & Q_k \end{bmatrix}.$$

Therefore, the characteristic polynomial $F(\mu,\lambda)$ of M is

$$F(\mu,\lambda)=\det\left(M-\lambda I_{Lp}\right)=\det\left(A^L-\mu\sum_{k=1}^{L}A^{L-k}B^kA^{k-1}-\lambda I_{Lp}+\mu^2C(\mu)\right).$$

The first p columns of $\left(A^L-\mu\sum_{k=1}^{L}A^{L-k}B^kA^{k-1}-\lambda I_{Lp}+\mu^2C(\mu)\right)$ are

$$\begin{bmatrix}(1-\lambda)I_p-\mu\left[LQ_L+(L-1)Q_1+\cdots+Q_{L-1}\right]+\mu^2C_{11}\\ I_p-\mu\left[(L-1)Q_L+(L-2)Q_1+\cdots+Q_{L-2}\right]+\mu^2C_{21}\\ I_p-\mu\left[(L-2)Q_L+(L-3)Q_1+\cdots+Q_{L-3}\right]+\mu^2C_{31}\\ \vdots\\ I_p-\mu\left(2Q_L+Q_1\right)+\mu^2C_{L-1\,1}\\ I_p-\mu Q_L+\mu^2C_{L1}\end{bmatrix},$$

the second p columns are

$$\begin{bmatrix}-(L-1)\mu Q_{L-1}+\mu^2C_{12}\\ -(L-1)\mu Q_{L-1}-\lambda I_p+\mu^2C_{22}\\ -(L-2)\mu Q_{L-1}+\mu^2C_{32}\\ \vdots\\ -2\mu Q_{L-1}+\mu^2C_{L-1\,2}\\ -\mu Q_{L-1}+\mu^2C_{L2}\end{bmatrix},$$

the next $(L-3)p$ columns are

$$\begin{bmatrix}-(L-2)\mu Q_{L-2}+\mu^2C_{13} & \dots & -2\mu Q_2+\mu^2C_{1\,L-1}\\ -(L-2)\mu Q_{L-2}+\mu^2C_{23} & \dots & -2\mu Q_2+\mu^2C_{2\,L-1}\\ -(L-2)\mu Q_{L-2}-\lambda I_p+\mu^2C_{33} & \dots & -2\mu Q_2+\mu^2C_{3\,L-1}\\ \vdots & & \vdots\\ -2\mu Q_{L-2}+\mu^2C_{L-1\,3} & \dots & -2\mu Q_2-\lambda I_p+\mu^2C_{L-1\,L-1}\\ -\mu Q_{L-2}+\mu^2C_{L3} & \dots & -\mu Q_2+\mu^2C_{L\,L-1}\end{bmatrix},$$

and the last p columns are

$$\begin{bmatrix}-\mu Q_1+\mu^2C_{1L}\\ -\mu Q_1+\mu^2C_{2L}\\ -\mu Q_1+\mu^2C_{3L}\\ \vdots\\ -\mu Q_1+\mu^2C_{L-1\,L}\\ -\mu Q_1-\lambda I_p+\mu^2C_{LL}\end{bmatrix},$$

where C_{ij}, $i,j=1,\dots,L$, are $p\times p$ matrices whose entries are polynomials in μ.

It is easy to see that if $\mu=0$, then $F(0,\lambda)=(-1)^{Lp}\lambda^{Lp-p}(\lambda-1)^p$. Hence, if $\mu=0$, we have an eigenvalue 0 of multiplicity $Lp-p$ and an eigenvalue 1 of multiplicity p. If μ is close enough to zero, the 0-eigenvalues will be close to the origin and therefore inside the unit circle. We need to prove that for sufficiently small positive μ, all the 1-eigenvalues will be inside the unit circle. Let $\lambda=\lambda(\mu)$ be a smooth function expressing the dependence of one of the 1-eigenvalues on μ. We will prove that $\frac{\mathrm{d}\lambda}{\mathrm{d}\mu}(0^+)<0$. It will be enough for our purposes, since it will show that the

trajectory $\lambda = \lambda(\mu)$ is entering the unit circle, and hence $\lambda(\mu)$ is inside the unit circle for sufficiently small positive μ.

By the definition of $\lambda(\mu)$, $\lambda(0+) = 1$ and $F(\mu, \lambda(\mu)) = 0$ for all μ. It follows that

$$\frac{\mathrm{d}^p F(\mu, \lambda(\mu))}{\mathrm{d}\mu^p} = 0. \tag{2.8}$$

To calculate the left-hand side of (2.8), we use the formula for the derivative of a determinant [23]. Note that substituting $\mu = 0$ and $\lambda = 1$ into each of the first p rows of the matrix $M - \lambda I_{Lp}$ leads to a row in which all of the entries are zeros, and therefore the determinant has a zero value. Therefore the only nonzero terms in $\frac{\mathrm{d}^p F(\mu,\lambda(\mu))}{\mathrm{d}\mu^p}$ after substituting $\mu = 0$ and $\lambda = 1$ (more precisely, taking $\mu \to 0^+$) are the terms with the first derivatives in the first p rows (there are $p!$ such terms). Hence taking the pth derivative is reduced to taking the first derivative of each of the first p rows. Substituting $\lambda = 1$ and $\mu \to 0^+$ we obtain

$$\frac{\mathrm{d}^p F(\mu, \lambda(\mu))}{\mathrm{d}\mu^p} = p! \det \begin{bmatrix} W_2 & W_3 \\ W_4 & -I_{(L-1)p\times(L-1)p} \end{bmatrix} = 0,$$

where $W_2 = -\lambda'(0^+) I_p - \sum_{k=0}^{L-1} (L-k) Q_{(k)_L}$,

$$W_3 = \begin{bmatrix} -(L-1)Q_{L-1} & -(L-2)Q_{L-2} & \dots & -2Q_2 & -Q_1 \end{bmatrix},$$

and $W_4 = [I_p\, I_p \dots I_p]^T$. Add all columns of $p \times p$ matrices to the first column of $p \times p$ matrices to obtain

$$\det \begin{bmatrix} W_5 & W_3 \\ 0_{(L-1)p\times p} & -I_{(L-1)p\times(L-1)p} \end{bmatrix} = 0,$$

where $W_5 = -\lambda'(0^+) I_p - L \sum_{k=1}^{L} Q_k$. Calculating the last determinant gives

$$\det \left[L \sum_{k=1}^{L} Q_k + \lambda'(0^+) I_p \right] = 0.$$

The last equation shows that $-\lambda'(0^+)$ is an eigenvalue of the matrix $L \sum_{k=1}^{L} Q_k$. Since $L \sum_{k=1}^{L} Q_k$ is positive definite, $-\lambda'(0^+) > 0$, and therefore $\lambda'(0^+) < 0$. This proves that for sufficiently small $\mu > 0$ the eigenvalues of the matrix M are strictly inside the unit circle, and hence the sequence x^k converges to x^*, and the convergence rate is linear. ∎

3. Initialization and distributed implementation. As mentioned in section 1, the IAG method is initiated with L points, $x^1, x^2, \dots, x^L$. Possible initialization strategies include setting $x^1 = x^2 = \cdots = x^L$ or generating the initial points using a single cycle of the standard incremental gradient method (1.5). Another possibility is the following. Given x^1, compute $d^1 = \nabla f_1(x^1)$. Then, for $1 \le k \le L-1$,

$$\begin{aligned} x^{k+1} &= x^k - \mu \frac{1}{k} d^k, \\ dkk + 1 &= d^k + \nabla f_{(k+1)_L}(x^{k+1}). \end{aligned} \tag{3.1}$$

Therefore, after $L-1$ iterations we obtain $x^1, \dots, x^L$ and $d^L = \sum_{l=1}^{L} \nabla f_l(x^l)$.

The key feature of the IAG method that makes it suitable for wireless sensor networks applications is that it can be implemented in a distributed manner. Consider a distributed system of L processors enumerated over $1, 2, \ldots, L$, each of which has access to one of the functions $f_l(x)$. The initialization (3.1) begins with x^1 at processor 1. Then processor 1 sets $d^1 = \nabla f_1(x^1)$ and transmits x^1 and d^1 to processor 2. Upon receiving x^{k-1} and d^{k-1} from processor $k-1$, processor k calculates x^k and d^k according to (3.1) and transmits them to processor $k+1$. The initialization phase is completed when processor L, upon receiving x^{L-1} and d^{L-1} from processor $L-1$, computes x^L and d^L according to (3.1) and transmits them to processor 1.

Once the initialization phase is completed, the algorithm progresses in a cyclic manner. Upon receiving x^{k-1} and d^{k-1} from processor $(k-1)_L$, processor $(k)_L$ computes x^k and d^k according to (1.2) and (1.3), respectively, and transmits them to processor $(k+1)_L$. Note that $\nabla f_{(k)_L}(x^{k-L})$ in (1.3) is available at processor $(k)_L$, since it was the last gradient computed at that processor. Therefore, the only gradient computation at processor $(k)_L$ is $\nabla f_{(k)_L}(x^k)$. At no phase of the algorithm do the processors share information regarding the complete function $f(x)$ or its gradient $\nabla f(x)$.

4. Application to wireless sensor networks. There are two motivations to use the IAG method: (a) reduced computational burden due to the evaluation of a single gradient per iteration compared to L gradients required for the steepest descent method; and (b) the possibility of a distributed implementation of the method in which each component has access to one of the functions $f_l(x)$. The second item has been shown to be very useful in the context of wireless sensor networks [38]. Wireless sensor networks provide means for efficient large scale monitoring of large areas [45]. Often the ultimate goal is to estimate certain parameters based on measurements that the sensors collect, giving rise to an optimization problem. If measurements from distinct sensors are modelled as statistically independent, the estimation problem takes the form of (1.1), where $f_l(x)$ is indexed by the measurements available at sensor l (see, e.g., [9, 33, 36, 37] and the references therein). When transmitting the complete set of data to a central processor is impractical due to bandwidth and power constraints, the IAG method can be implemented in a distributed manner as described in section 3. In the following sections we consider two such estimation problems.

4.1. Robust estimation. One of the benefits of a wireless sensor network is the ability to deploy a large number of low cost sensors to densely monitor a certain area [45]. Because low cost sensors have limited reliability, the system must be designed to be robust to the possibility of individual sensor failures. In estimation tasks, this means that some of the sensors will contribute unreliable measurements, namely outliers. In [36] the authors suggest the use of robust statistics to alleviate the influence of outliers in the data (see [20] or, specifically in the context of optimization, [34, p. 347]). The robust statistics framework uses objective functions that give less weight to outliers. A common objective function used to this end is the function "Fair" [39, p. 110], given by

$$g(x) = c^2 \left[\frac{|x|}{c} - \log\left(1 + \frac{|x|}{c}\right) \right]. \tag{4.1}$$

Following [36] we simulate a sensor network for measuring pollution levels and assume that a certain percentage of the sensors are damaged and provide unreliable measurements. Each sensor collects a single noisy measurement of the pollution level,

and the estimate of the average pollution level is found by minimizing the objective function defined by

$$f(x) = \sum_{l=1}^{L} f_l(x), \tag{4.2}$$

where $x \in \mathbb{R}$, and

$$f_l(x) = \frac{1}{L} g(x - y_l),$$

where y_l is the measurement collected by sensor l. There were $L = 50$ sensors in the simulation. To reflect the possibility of faulty sensors, half of the samples were generated according to a Gaussian distribution with mean $m_1 = 10$ and unit variance ($\sigma_1^2 = 1$), and the other half were generated according to a Gaussian distribution with mean $m_2 = 10$ and ten times higher variance ($\sigma_2^2 = 10$). The coefficient c in (4.1) was chosen to be 10.

For positive x, the first derivative of $g(x)$ is $\frac{x}{1+x/c}$, and for negative x it is $\frac{x}{1-x/c}$. Hence, $g'(0+) = g'(0-) = 0$. The continuity of $g(x)$ implies then that it is differentiable at zero despite the term $|x|$. Therefore, the first derivative of $g(x)$ is $\frac{x}{1+|x|/c}$, it is continuous, and it is bounded by c. Considering positive and negative x's separately also shows that $g''(0+) = g''(0-) = 1$ and that, in general, the second derivative of $g(x)$ is $\frac{1}{(1+|x|/c)^2}$, which is bounded by 1. Hence both Assumptions 1 and 2 hold. In addition, since $\frac{1}{(1+|x|/c)^2}$ is strictly positive, $g(x)$ is strictly convex, and therefore $f(x)$ is strictly convex as well. Since both $\lim_{x\to\infty} f(x)$ and $\lim_{x\to-\infty} f(x)$ diverge to ∞, $f(x)$ has no directions of recession, and therefore, by Theorem 27.1(d) in [41, p. 265], the minimum set of $f(x)$ is nonempty. The minimum is unique by the strict convexity of $f(x)$. Since $g''(x)$ is continuous and positive everywhere, Assumption 3 is satisfied. The strict convexity of $f(x)$ implies that Assumption 4 holds as well (see Remark 2).

Both the standard incremental gradient method (1.5) with a constant step size $\mu(k) = \mu$ (abbreviated as IG in the figures) and the IAG method with the initialization (3.1) were implemented with several choices of step size μ. The initial point x^1 was set to 0. In Figure 4.1 the trajectories of the two methods are presented. The solid straight line corresponds to the minimum point x^*. It is seen that when the step size is sufficiently small, IAG increases more rapidly towards x^* than the standard incremental gradient in the early iterations. Furthermore, as predicted by the theory, IAG converges to the true limit, whereas the incremental gradient method converges to a limit cycle. For a larger step size the IAG method overshoots due to its heavy ball characteristic (1.4). When the step size is too large, the IAG method no longer converges, but the incremental gradient method still converges to a limit cycle. We have observed this behavior for other values of the parameters $m_1, m_2, \sigma_1^2, \sigma_2^2, c$ as well.

We also compared the IAG method with the incremental gradient method with a diminishing step size, with Bertsekas' hybrid method [5], and with Tseng's incremental gradient with momentum [46] in terms of number of iterations to convergence. To optimize the performance of the incremental gradient method with a diminishing step size, a relatively large constant step size $\mu = 0.2$ is used until convergence to a limit cycle is detected, and then the diminishing step size is $\mu(k) = .2\mu/(\widetilde{k} - k)$, where $\widetilde{k}$ is the first iteration in which a limit cycle is detected. Convergence to a limit cycle is declared when $|x^k - x^{k-L}| < .01$ for k a multiple of L. To describe the parameters

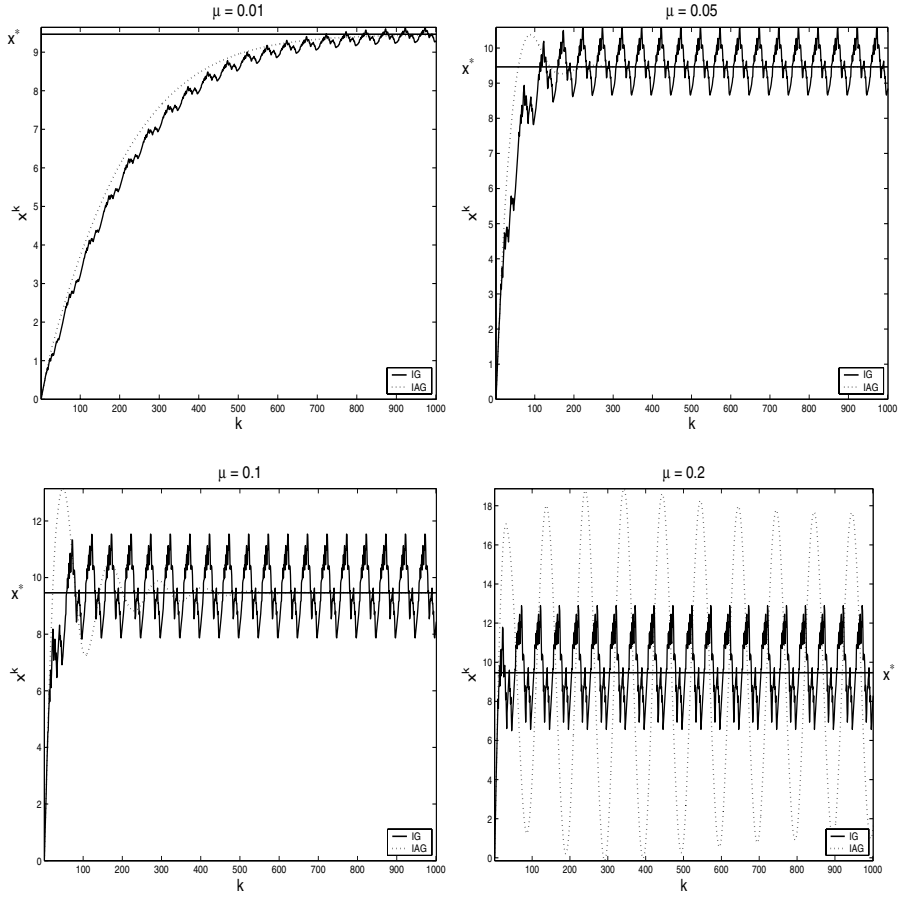


FIG. 4.1. *Trajectories taken by the IG and IAG methods for the robust "Fair" estimation problem.*

used in the hybrid method, we switch to the notation in [5]. We set $\gamma = 0.05$ and $\alpha(\mu)$ as defined in equation (47) in [5], with $\phi(\mu) = \zeta(1-\mu)$, where $\zeta = 2.5$. The transition parameter μ is kept at zero; i.e., the iterates are identical to the incremental gradient method until convergence to a limit cycle is detected as described above. Once a limit cycle is detected, μ is updated after every cycle according to $\mu := 1.5\mu + 0.3$, i.e., $\widehat{n} = 1$. These parameters seemed to optimize the performance of the hybrid method. The parameters of the incremental gradient with momentum term were set according to the recommendation in [46], which seemed to optimize the performance of the method in our application as well. In particular, we set $\epsilon_0 = 1$, $\epsilon_1 = \epsilon_2 = 0.00001$, $\epsilon_3 = 1000$, $\eta = 1.5 f(x_1^0) + 100$, $\rho = \infty$, $\omega = 0.5$, $\zeta = 0.8$, and $\lambda_1 + \lambda_2 + \cdots + \lambda_m = 1$. For the IAG method we set $\mu = 0.05$. The convergence point was specified to be the first iteration for which all subsequent iterations satisfy $|x^k - x^*| < \epsilon$. Since the IAG and the hybrid methods outperform the incremental gradient method with a diminishing step size and the incremental gradient with momentum term by a large margin, ϵ was specified to be 0.01 for the IAG and the hybrid method and 0.1 for the incremental gradient method with a diminishing step size and the incremental gradient with momentum term. The average number of iterations until convergence and its standard deviation were estimated from 100 Monte Carlo simulations and are summarized in Table 4.1.

TABLE 4.1
Number of iterations to convergence.

	IAG $\epsilon = 0.01$	Hybrid $\epsilon = 0.01$	IG diminishing step size $\epsilon = 0.1$	IG momentum term $\epsilon = 0.1$
Mean	290	589	601	2063
Std	23	135	258	919

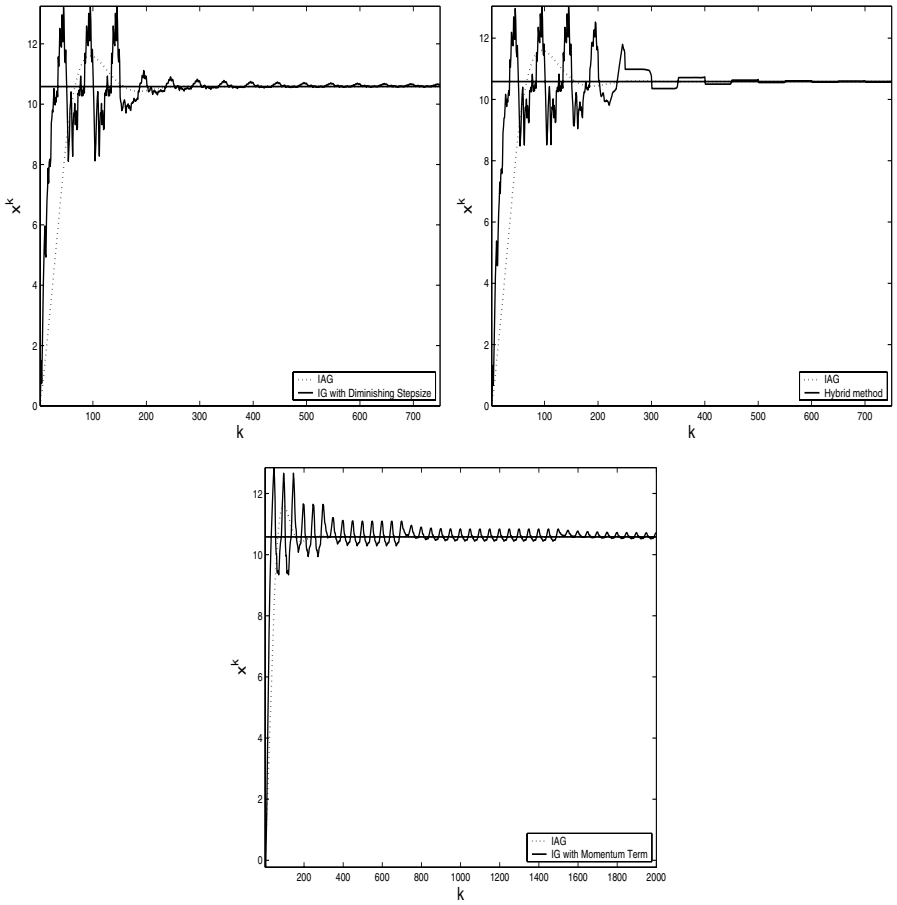


FIG. 4.2. *IAG compared to IG with diminishing step size, to the hybrid method, and to IG with momentum term.*

The trajectory taken by the different methods in one of these simulations is presented in Figure 4.2. It is seen that for this application, the IAG method performs best. Further experimentation is required to make more general conclusions.

4.2. Source localization. This section presents a simulation of a sensor network for localizing a source that emits acoustic waves. L sensors are distributed on the perimeter of a field at known spatial locations, denoted r_l, $l = 1, \ldots, L$, where $r_l \in \mathbb{R}^2$. Each sensor collects a noisy measurement of the acoustic signal transmitted by the source, denoted y_l, at an unknown location x. Based on a far-field assumption and an isotropic acoustic wave propagation model [13, 24, 36, 42, 43], the problem of estimation of source location can be formulated as a nonlinear least squares problem.

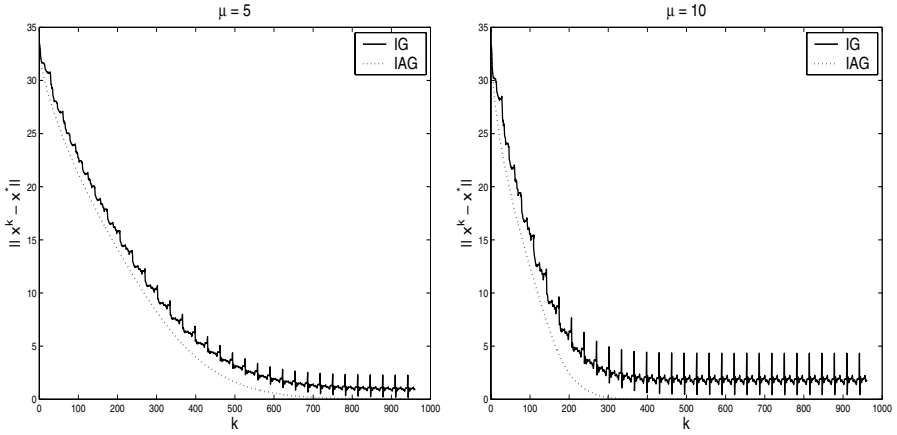


FIG. 4.3. *Distance of IG and IAG iterates to the optimal solution x^* for source localization problem.*

The objective function is again of the form (4.2), but now

$$f_l(x) = \left(y_l - g(||r_l - x||^2)\right)^2, \tag{4.3}$$

$x \in \mathbb{R}^2$, and

$$g(z) = \begin{cases} A/z & : \quad z \geq A/\epsilon, \\ 2\epsilon - \epsilon^2 z/A & : \quad z < A/\epsilon. \end{cases} \tag{4.4}$$

In (4.3) $g(\cdot)$ models the received signal strength as a function of the squared distance. In (4.4) A is a known constant characterizing the source's signal strength. For $z \geq A/\epsilon$ (far-field source), the source's signal strength has isotropic attenuation as an inverse function of the squared distance, while for $z < A/\epsilon$ (near-field source), the attenuation is linear in the squared distance. It is easy to see that Assumptions 1 and 2 are satisfied, and therefore Proposition 2.4 holds. Clearly, since $f(x)$ is multimodal in this case, Assumptions 3 and 4 cannot hold. However, it was observed in our experiments that when the source is sufficiently distant from the sensors, the objective function has a single minimum inside the observed field (see Figure 4.4 for a contour plot of the objective function) and, when initiated not too far from the minimum point, the IAG method has good convergence properties. This suggests the possible application of the IAG method under weaker assumptions than those considered in this paper and motivates further investigation into its properties.

In the numerical experiment, $L = 32$ sensors are distributed equidistantly on the perimeter of a 100×100 field. The source is located at the point $[60, 60]$ and emits a signal with strength $A = 1000$. The sensors' noisy measurements were generated according to a Gaussian distribution with a mean equal to the true signal power and unit variance. Both the incremental gradient method with a constant step size and the IAG method with the initialization (3.1) were initiated at the point $[40, 40]$. The error term $||x^k - x^*||$ as a function of the iteration number is presented in Figure 4.3 for two choices of step size. The actual path taken by the methods for step size $\mu = 10$ is presented in Figure 4.4, where the asterisk denotes the true minimum point of the objective function. It is seen that, as the theory predicts, the incremental

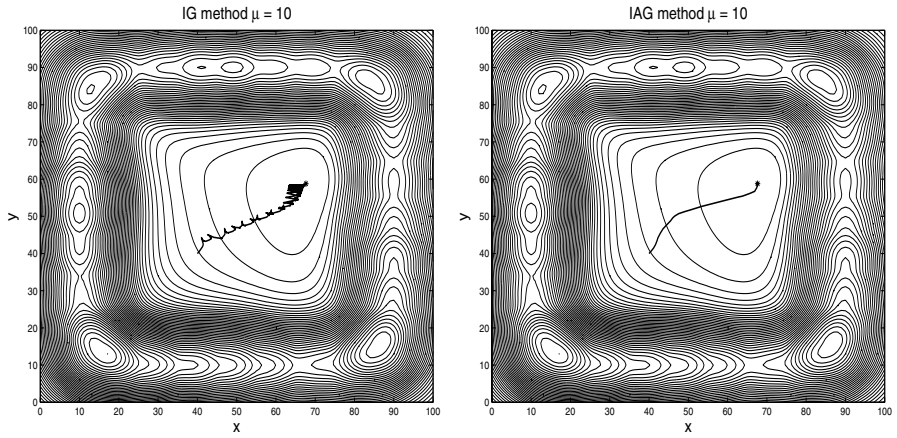


FIG. 4.4. *Path taken by the IG and IAG methods for source localization problem.*

gradient method exhibits oscillations near the eventual limit, whereas the IAG method converges to the minimum. In this scenario, the IAG method outperforms the IG method at early iterations as well.

REFERENCES

[1] S. AHN, J. A. FESSLER, D. BLATT, AND A. O. HERO, *Convergent incremental optimization transfer algorithms: Application to tomography*, IEEE Trans. Med. Imag., 25 (2006), pp. 283–296.
[2] B. M. BELL, *The iterated Kalman smoother as a Gauss–Newton method*, SIAM J. Optim., 4 (1994), pp. 626–636.
[3] A. BEN-TAL, T. MARGALIT, AND A. NEMIROVSKI, *The ordered subsets mirror descent optimization method with applications to tomography*, SIAM J. Optim., 12 (2001), pp. 79–108.
[4] D. P. BERTSEKAS, *Incremental least squares methods and the extended Kalman filter*, SIAM J. Optim., 6 (1996), pp. 807–822.
[5] D. P. BERTSEKAS, *A new class of incremental gradient methods for least squares problems*, SIAM J. Optim., 7 (1997), pp. 913–926.
[6] D. P. BERTSEKAS, *Nonlinear Programming*, 2nd ed., Athena Scientific, Belmont, MA, 1999.
[7] D. P. BERTSEKAS AND J. N. TSITSIKLIS, *Neuro-Dynamic Programming*, Athena Scientific, Belmont, MA, 1996.
[8] D. P. BERTSEKAS AND J. N. TSITSIKLIS, *Gradient convergence in gradient methods with errors*, SIAM J. Optim., 10 (2000), pp. 627–642.
[9] D. BLATT AND A. HERO, *Distributed maximum likelihood for sensor networks*, in Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, Canada, 2004, pp. 929–932.
[10] C. BYRNE, *Choosing parameters in block-iterative or ordered subset reconstruction algorithms*, IEEE Trans. Image Process., 14 (2005), pp. 321–327.
[11] Y. CENSOR AND G. T. HERMAN, *Block-iterative algorithms with underrelaxed Bregman projections*, SIAM J. Optim., 13 (2002), pp. 283–297.
[12] Y. CENSOR, A. R. D. PIERRO, AND M. ZAKNOON, *Steered sequential projections for the inconsistent convex feasibility problem*, Nonlinear Anal., 59 (2004), pp. 385–405.
[13] J. C. CHEN, K. YAO, AND R. E. HUDSON, *Source localization and beamforming*, IEEE Signal Process. Mag., 19 (2002), pp. 30–39.
[14] M. COLLINS, R. E. SCHAPIRE, AND Y. SINGER, *Logistic regression, AdaBoost and Bregman distances*, Mach. Learn., 48 (2002), pp. 253–285.
[15] W. C. DAVIDON, *New least-square algorithms*, J. Optim. Theory Appl., 18 (1976), pp. 187–197.
[16] J. FRIEDMAN, T. HASTIE, AND R. TIBSHIRANI, *Additive logistic regression: A statistical view of Boosting*, Ann. Statist., 38 (2000), pp. 337–374.
[17] A. A. GAIVORONSKI, *Convergence analysis of parallel backpropagation algorithm for neural*

networks, Optim. Methods Softw., 4 (1994), pp. 117–134.
[18] L. Grippo, *A class of unconstrained minimization methods for neural networks training*, Optim. Methods Softw., 4 (1994), pp. 135–150.
[19] L. Grippo, *Convergent on-line algorithms for supervised learning in neural networks*, IEEE Trans. Neural Networks, 11 (2000), pp. 1284–1299.
[20] P. Huber, *Robust Statistics*, John Wiley and Sons, New York, 1981.
[21] V. M. Kibardin, *Decomposition into functions in the minimization problem*, Automat. Remote Control, 40 (1980), pp. 1311–1321.
[22] K. C. Kiwiel, *Convergence of approximate and incremental subgradient methods for convex optimization*, SIAM J. Optim., 14 (2004), pp. 807–840.
[23] E. Kreyszic, *Advanced Engineering Mathematics*, John Wiley and Sons, New York, 1988.
[24] D. Li and Y. H. Hu, *Energy-based collaborative source localization using acoustic microsensor array*, EURASIP J. Appl. Signal Process., (2003), pp. 321–337.
[25] Z. Q. Luo, *On the convergence of the LMS algorithm with adaptive learning rate for linear feedforward networks*, Neural Comput., 3 (1991), pp. 226–245.
[26] Z. Q. Luo and P. Tseng, *Analysis of an approximate gradient projection method with applications to the backpropagation algorithm*, Optim. Methods Softw., 4 (1994), pp. 85–101.
[27] O. L. Mangasarian, *Mathematical programming in neural networks*, ORSA J. Comput., 5 (1993), pp. 349–360.
[28] O. L. Mangasarian and M. V. Solodov, *Serial and parallel backpropagation convergence via nonmonotone perturbed minimization*, Optim. Methods Softw., 4 (1994), pp. 103–116.
[29] R. Meyer and C. Burrus, *A unified analysis of multirate and periodically time-varying digital filters*, IEEE Trans. Circuits and Systems, 22 (1975), pp. 162–168.
[30] H. Moriyama, N. Yamashita, and M. Fukushima, *The incremental Gauss-Newton algorithm with adaptive stepsize rule*, Comput. Optim. Appl., 26 (2003), pp. 107–141.
[31] R. M. Neal and G. E. Hinton, *A view of the EM algorithm that justifies incremental, sparse, and other variants*, in Learning in Graphical Models, M. I. Jordan, ed., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1994, pp. 355–368.
[32] A. Nedić and D. P. Bertsekas, *Incremental subgradient methods for nondifferentiable optimization*, SIAM J. Optim., 12 (2001), pp. 109–138.
[33] R. D. Nowak, *Distributed EM algorithms for density estimation and clustering in sensor networks*, IEEE Trans. Signal Process., 51 (2003), pp. 2245–2253.
[34] B. T. Polyak, *Introduction to Optimization*, Optimization Software, New York, 1987.
[35] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing*, Prentice–Hall, Englewood Cliffs, NJ, 1996.
[36] M. G. Rabbat and R. D. Nowak, *Decentralized source localization and tracking*, in Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, Canada, 2004, pp. 921–924.
[37] M. G. Rabbat and R. D. Nowak, *Distributed optimization in sensor networks*, in Proceedings of the Third International Symposium on Information Processing in Sensor Networks, Berkeley, CA, 2004, ACM Press, New York, 2004, pp. 20–27.
[38] M. G. Rabbat and R. D. Nowak, *Quantized incremental algorithms for distributed optimization*, IEEE J. Selected Areas in Communications, 23 (2005), pp. 798–808.
[39] W. J. J. Rey, *Introduction to Robust and Quasi-Robust Statistical Methods*, Springer-Verlag, Berlin, 1983.
[40] A. W. Roberts and D. E. Varberg, *Convex Functions*, Academic Press, New York, 1973.
[41] R. T. Rockafeller, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[42] X. Sheng and Y. H. Hu, *Energy based acoustic source localization*, in Proceedings of the Second International Conference on Information Processing in Sensor Networks, Palo Alto, CA, 2003, Lecture Notes in Comput. Sci. 2634, Z. Feng and G. Leonidas, eds., Springer-Verlag, New York, 2003, pp. 285–300.
[43] X. Sheng and Y. H. Hu, *Maximum likelihood multiple-source localization using acoustic energy measurements with wireless sensor networks*, IEEE Trans. Signal Process., 53 (2005), pp. 44–53.
[44] M. V. Solodov, *Incremental gradient algorithms with stepsizes bounded away from zero*, Comput. Optim. Appl., 11 (1998), pp. 23–35.
[45] R. Szewczyk, E. Osterweil, J. Polastre, M. Hamilton, A. Mainwaring, and D. Estrin, *Habitat monitoring with sensor networks*, Comm. ACM, 47 (2004), pp. 34–40.
[46] P. Tseng, *An incremental gradient(-projection) method with momentum term and adaptive stepsize rule*, SIAM J. Optim., 8 (1998), pp. 506–531.

SIAM J. OPTIM.
Vol. 18, No. 1, pp. 52–64

ON THE RELATIONSHIP BETWEEN THE CONVERGENCE RATES OF ITERATIVE AND CONTINUOUS PROCESSES*

RAPHAEL HAUSER† AND JELENA NEDIĆ†

Abstract. Considering iterative sequences that arise when approximate solutions x_k to a numerical problem are updated by $x_{k+1} = x_k + v(x_k)$, where v is a differentiable vector field, we derive necessary and sufficient conditions for such discrete processes to converge to a stationary point of v at different Q-rates in terms of a similar notion of fast convergence for the corresponding continuous processes.

Key words. Q-convergence, superlinear convergence, nonlinear analysis, unconstrained optimization, root finding

AMS subject classifications. Primary, 34A34, 41A25, 37N40; Secondary, 34G20, 90C26

DOI. 10.1137/040620631

1. Introduction.

In this paper we study sequences $(x_k)_{\mathbb{N}_0}$ which, given a starting point x_0, consist of points that are iteratively related to one another via the rule

$$x_{k+1} = x_k + v(x_k), \tag{1.1}$$

where $v : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ is a differentiable vector field with Jacobian $J = \mathrm{D}v$, and where D is a convex open domain. We are interested in the situation where x_k converges to a stationary point x^* of v, i.e., a point where $v(x^*) = 0$. Sequences of the form (1.1) appear in many areas of numerical analysis where an approximate solution x_k is iteratively improved, notably in unconstrained optimization and in zero-finding problems. One of the major objectives in designing iterative schemes of this kind is to ensure that they converge at a provably fast rate to a point x^* which represents a solution of interest. The concept of fast convergence used in this paper is that of *uniform* Q-convergence. Related but not entirely equivalent notions of Q-convergence have been discussed in the literature; see, e.g., [5].

DEFINITION 1.1. *We say that the process* (1.1) *converges to* x^* *uniformly at Q-convergence rate* $q > 1$ *in the ball* $\mathrm{B}_\rho(x^*)$ *if there exists* $\beta > 0$ *such that*

$$\|x + v(x) - x^*\| \le \beta \|x - x^*\|^q \tag{1.2}$$

for all $x \in \mathrm{B}_\rho(x^*)$*, where* $\|\cdot\|$ *denotes the Euclidean norm.*

Setting $\tilde{\rho} := \min(\rho, \beta^{\frac{1}{q-1}})$, another way to express this is that if the sequence $(x_k)_{\mathbb{N}}$ ever enters the ball $\mathrm{B}_{\tilde{\rho}}(x^*)$, then it converges to x^*, and each iteration starting from within the ball improves the accuracy of x_k as an approximation of x^* to about q times as many correct digits as beforehand. The constant β is called the *convergence factor*. For example, the well-known Kantorovich theorem [3] shows that Newton's method

*Received by the editors December 10, 2004; accepted for publication (in revised form) August 8, 2006; published electronically February 2, 2007.
http://www.siam.org/journals/siopt/18-1/62063.html
†Oxford University Computing Laboratory, Wolfson Building, Parks Road, Oxford, OX1 3QD, UK (hauser@comlab.ox.ac.uk, jelena.nedic@googlemail.com). The first author was supported through grant GR/S34472 from the Engineering and Physical Sciences Research Council of the UK. The second author was supported through the Clarendon Fund, Oxford University Press and ORS Award, Universities UK.

for zero-solving converges Q-quadratically under standard regularity assumptions. A weaker notion of fast convergence is the following.

DEFINITION 1.2. *We say that the process* (1.1) *converges uniformly Q-superlinearly to* x^* *if*

$$\lim_{x \to x^*} \frac{\|x + v(x) - x^*\|}{\|x - x^*\|} = 0. \tag{1.3}$$

That is, asymptotically, each iteration adds more than any fixed number of additional correct digits to the current approximation of x^*.

Let us briefly comment on the notation used in this paper. All identity matrices are denoted by I, irrespective of their dimension. We write S^{n-1} for the set of unit vectors $\{x \in \mathbb{R}^n : \|x\| = 1\}$ in $\mathbb{R}^n$. This is a standard notation that accounts for the fact that the sphere S^{n-1} is a $(n-1)$-dimensional manifold. Another standard notation, already used, is to write $B_\delta(x) := \{y \in \mathbb{R}^n : \|y - x\| < \delta\}$ for the open Euclidean ball of radius δ in $\mathbb{R}^n$. We denote inner products by $\langle \cdot, \cdot \rangle$ and use $\cdot$ for scalar multiplication instead where it helps improve the readability of the text. If x is a nonzero vector in $\mathbb{R}^n$, we write $\mathfrak{n}(x) := \|x\|^{-1} \cdot x$ for its normalization.

1.1. Overview. The discrete dynamical system (1.1) has a continuous analogue obtained when damping with an infinitesimal step size is applied. The associated flow is defined by

$$\frac{\partial}{\partial t}\varphi(x,t) = v(\varphi(x,t)), \qquad \varphi(x,0) = x. \tag{1.4}$$

In other words, in the continuous process one chooses a starting point x and follows the flux line $t \mapsto \varphi(x,t)$, obtained by integrating the ODE (1.4) to a limit point $x^* = \lim_{t\to\infty} \varphi(x,t)$. Note that the *flow-conservation property*

$$\varphi(x, t+\tau) = \varphi\big(\varphi(x,t), \tau\big) \tag{1.5}$$

holds for all x, t, and τ for which both sides are well defined.

The main goal of this paper is to investigate the notion of Q-convergence for the discrete process (1.1) via a new notion of fast convergence for the associated continuous process (1.4). This is done in section 3, where we define the notions of exponential and p-exponential convergence for (1.4) and show that they are equivalent to uniform Q-superlinear and Q-convergence of rate $p+1$ for (1.1); see Theorem 3.4.

Our results shed new light on the well-established notion of Q-convergence, as exponential convergence has an easy geometric interpretation in terms of the flux lines of (1.4): not only do these have to converge to x^* exponentially fast in t, but also does a rotational component have to die out sufficiently quickly.

It is often observed that continuous dynamical systems are easier to analyze than discrete ones. Useful applications of Theorem 3.4 as an analytic tool derive from this observation. We illustrate this in section 5, where we discuss an example of a vector field $v(x)$ with the property that none of the rescaled vector fields $\lambda(x)v(x)$ leads to a discrete process (1.1) with a faster Q-convergence rate than the process corresponding to $v(x)$. In this context $\lambda(x)$ can be chosen as an arbitrary positive differentiable scalar function. This observation is relevant in unconstrained optimization, as it shows that a line search cannot be expected to speed up the asymptotic convergence rate of a search direction field.

Continuous methods for solving zero-finding and optimization problems have been proposed by various authors; see, e.g., [1, 2] for relevant ideas and references. To avoid

any confusion, we point out that our paper is not intended as a direct contribution to this discussion. Instead, our focus is on understanding convergence rates alone and on deriving a useful observation in the context of unconstrained optimization.

To prepare the analysis of our main result in section 3, an exact characterization of the vector fields that make the process (1.1) converge fast under either notion of uniform Q-convergence is given in section 2: Theorem 2.3 shows that when the Jacobian $J(x)$ is sufficiently smooth at x^*, then the process (1.1) converges at a quantifiable Q-convergence rate if and only if $v(x^*) = 0$ and $J(x^*) = -\mathrm{I}$. This result is also interesting in its own right.

2. Fast convergence of the discrete process. In this section we consider the discrete dynamical system (1.1). We will see that uniform Q-superlinear convergence and uniform Q-convergence at a given rate are characterized by the differential properties of $v(x)$ in a neighborhood of x^*.

Let $p > 0$. Recall that the tensor field $J(x)$ is said to be *p-Hölder continuous at* x^* if there exist constants $\alpha > 0$ and $\varrho > 0$ such that

$$\|J(x) - J(x^*)\| \leq \alpha \|x - x^*\|^p \tag{2.1}$$

for all $x \in \mathrm{B}_\varrho(x^*)$.

LEMMA 2.1. *Let J be continuous at x^*, $v(x^*) = 0$, and $J(x^*) = -\mathrm{I}$. Then* (1.1) *converges uniformly Q-superlinearly to x^*. If J is furthermore p-Hölder continuous at x^* then* (1.1) *converges uniformly at the Q-convergence rate $p+1$ in a neighborhood of x^*.*

Proof. Using $v(x^*) = 0$, we find that for all $x \in D$,

$$\begin{aligned}\lim_{x\to x^*} \frac{\|x + v(x) - x^*\|}{\|x - x^*\|} &= \lim_{x\to x^*} \frac{\left\| \left(\mathrm{I} + \int_0^1 J(x^* + t(x - x^*))dt \right)(x - x^*) \right\|}{\|x - x^*\|} \\ &\leq \lim_{x\to x^*} \left\| \mathrm{I} + \int_0^1 J(x^* + t(x - x^*))dt \right\| = 0,\end{aligned}$$

where the last equality follows from the continuity of J at x^*. This shows the first claim. The second claim is established as follows:

$$\begin{aligned}\|x + v(x) - x^*\| &= \left\| \left(\mathrm{I} + \int_0^1 J(x^* + t(x - x^*))dt \right)(x - x^*) \right\| \\ &\leq \left\| \mathrm{I} + \int_0^1 J(x^* + t(x - x^*))dt \right\| \cdot \|x - x^*\| \\ &\leq \int_0^1 \alpha t^p \|x - x^*\|^p dt \cdot \|x - x^*\| = \frac{\alpha}{p+1} \cdot \|x - x^*\|^{p+1}. \qquad \square\end{aligned}$$

We remark that Lemma 2.1 is a minor adaptation of a special case of Theorem 10.1.6 [4] by Ortega and Rheinboldt, which says that if $G : D \subset \mathbb{R}^n \to \mathbb{R}^n$ has a fixed point $x^* \in D$ at which it is F-differentiable with $G'(x^*) = 0$, then x^* is a point of attraction of the process $x_{k+1} := G(x_k)$, and furthermore, if $G(x)$ is p-Hölder continuous at x^*, then the process converges at order p. In the context of Lemma 2.1 one can choose $G(x) = x + v(x)$, so that the criterion $G'(x^*) = 0$ becomes $\mathrm{I} + J(x^*) = 0$ and the fixed-point criterion $G(x^*) = x^*$ becomes $x^* = x^* + v(x^*)$. Furthermore, p-Hölder continuity of $G(x)$ at x^* means the existence of a constant $\beta > 0$ such that

$$\|x + v(x) - x^*\| = \|G(x) - G(x^*)\| \leq \beta \|x - x^*\|^p \tag{2.2}$$

in a neighborhood of x^*, which is in fact nothing but the definition of uniform Q-convergence of order p; see Definition 1.1. Thus, Theorem 10.1.6 in [4] applies to more general maps under weaker differentiability assumptions than Lemma 2.1, but by making the assumption (2.2), order-p convergence is a forgone conclusion in the specific case that is of interest here.

We are now going to prove a partial inverse of Lemma 2.1.

LEMMA 2.2. *If* (1.1) *converges uniformly Q-superlinearly to* x^**, then* $v(x^*) = 0$ *and* $J(x^*) = -\mathrm{I}$.

Proof. Equation (1.3) implies that for all $\varepsilon > 0$ there exists $\delta_\varepsilon > 0$ such that $x \in \mathrm{B}_{\delta_\varepsilon}(x^*)$ implies

$$\|x + v(x) - x^*\| \le \varepsilon \|x - x^*\|. \tag{2.3}$$

Taking limits $x \to x^*$ on both sides and using the continuity of v, we obtain $\|v(x^*)\| \le 0$, which shows that $v(x^*) = 0$. Next, let $z \in \mathrm{S}^{n-1}$ and consider the sequence $(x_n)_{\mathbb{N}}$ defined by $x_n = x^* + z/n$. Then (1.3) and $v(x^*) = 0$ show that

$$\|\big(\mathrm{I} + J(x^*)\big)z\| = \lim_{n\to\infty} \frac{\|x_n - x^* + v(x_n) - v(x^*)\|}{\|x_n - x^*\|} = 0.$$

But this shows that $J(x^*)z = -z$, and since z was an arbitrary unit vector, it follows that $J(x^*) = -\mathrm{I}$. ∎

It thus emerges that if continuity or Hölder continuity of J at x^* is given, then the two preceding lemmas yield the following exact characterization of fast convergence.

THEOREM 2.3. *If* J *is continuous at* x^**, then* (1.1) *converges to* x^* *uniformly Q-superlinearly if and only if* $v(x^*) = 0$ *and* $J(x^*) = -\mathrm{I}$. *If* J *is p-Hölder continuous at* x^**, then* (1.1) *converges uniformly at the Q-convergence rate* $p+1$ *in a neighborhood of* x^* *if and only if* $v(x^*) = 0$ *and* $J(x^*) = -\mathrm{I}$.

Proof. Observe that Q-convergence at a rate $p + 1 > 1$ implies Q-superlinear convergence. Everything else follows directly from Lemmas 2.1 and 2.2. ∎

Note that Theorem 2.3 shows that the only difference between Q-superlinear convergence and convergence at Q-rates $p + 1 > 1$ consists in the smoothness of the Jacobian of v in a neighborhood of x^*.

3. Fast convergence of the continuous process. In this section we introduce notions of fast convergence for the continuous dynamical system (1.4) which we will show to be equivalent to uniform Q-superlinear convergence and uniform Q-convergence of rate $p + 1$, respectively, of the discrete system (1.1).

The following property of differential inequalities is well known; see, e.g., [6].

LEMMA 3.1. *Let* $\frac{d}{dt}y(t) = g(t, y(t))$, $y(0) = y_0$, $t \ge 0$, *where* $g \in C(\mathbb{R}_+, \mathbb{R})$ *and* $y \in C^1(\mathbb{R}_+, \mathbb{R})$, *and let* $\frac{d}{dt}z(t) \lessgtr g(t, z(t))$, $z(0) = y_0$, $t \ge 0$. *If* $g(t, x)$ *is monotone increasing in* x *or of the form* $g(t, x) = h(t)$ *or* $g(t, x) = h(t)x$, *then* $z(t) \lessgtr y(t)$ *for all* $t \ge 0$.

We will be interested in the normalized vector

$$\mathfrak{n}(\varphi(x,t) - x^*) := \frac{\varphi(x,t) - x^*}{\|\varphi(x,t) - x^*\|} \in \mathrm{S}^{n-1} := \big\{x \in \mathbb{R}^n : \|x\| = 1\big\}$$

and the speed $\|\frac{\partial}{\partial t}\mathfrak{n}(\varphi(x,t) - x^*)\|$ which we will call the *angular speed* and which can be interpreted as the absolute angle traversed by the flux line $\varphi(x,t)$ with respect to x^* per unit time.

LEMMA 3.2. *For $\varphi(x,t)$ as defined by (1.4) it is true that*

$$\frac{\partial}{\partial t}\mathfrak{n}(\varphi(x,t)-x^*) = \frac{(\mathrm{I}-P(x,t))\,v(\varphi(x,t))}{\|\varphi(x,t)-x^*\|}, \tag{3.1}$$

where $P(x,t)$ denotes the orthogonal projection of $\mathbb{R}^n$ onto $\operatorname{span}\{\varphi(x,t)-x^*\}$.

Proof. The proof is a straightforward calculation. ∎

DEFINITION 3.3. *Let $v : D \to \mathbb{R}$ be a differentiable vector field with Jacobian J and let $x^* \in D$ be a stationary point of v, that is, $v(x^*) = 0$. We say that the continuous dynamical system (1.4) defined by v converges* exponentially *to x^* if for all $\varepsilon \in (0,1)$ there exists $\rho_\varepsilon > 0$ such that $x \in \mathrm{B}_{\rho_\varepsilon}(x^*) \setminus \{x^*\}$ and $t \geq 0$ imply*

$$\mathrm{e}^{-(1+\varepsilon)t}\,\|x-x^*\| \leq \|\varphi(x,t)-x^*\| \leq \mathrm{e}^{-(1-\varepsilon)t}\,\|x-x^*\|, \tag{3.2}$$

$$\left\|\frac{\partial}{\partial t}\mathfrak{n}(\varphi(x,t)-x^*)\right\| \leq \varepsilon. \tag{3.3}$$

We say that (1.4) converges p-exponentially to x^* if (3.2) holds and there exist constants $\xi > 0$ and $\gamma > 0$ such that $\rho_\varepsilon > \xi\varepsilon^{1/p}$ for all ε small enough and

$$\left\|\frac{\partial}{\partial t}\mathfrak{n}(\varphi(x,t)-x^*)\right\| \leq \gamma\,\mathrm{e}^{-(1-\varepsilon)pt}\,\|x-x^*\|^p \tag{3.4}$$

for all $x \in \mathrm{B}_{\rho_\varepsilon}(x^*)$ and $t \geq 0$.

Our goal now is to establish the following equivalence that constitutes the main result of this paper. Note that in contrast to Theorem 2.3, this theorem does not make any assumptions on the Hölder continuity of J at x^*.

THEOREM 3.4. *The notion of exponential convergence of the continuous dynamical system (1.4) is equivalent to the notion of uniform Q-superlinear convergence of the discrete system (1.1). Likewise, the notion of p-exponential convergence of (1.4) is equivalent to the notion of uniform Q-convergence of rate $p+1$.*

Proof. The proof follows immediately from Lemmas 3.5 and 3.6. ∎

LEMMA 3.5. *If (1.1) converges uniformly Q-superlinearly to x^*, then (1.4) converges exponentially to x^*. Moreover, if (1.1) converges uniformly at Q-convergence rate $p+1 > 1$, then (1.4) converges p-exponentially.*

Proof. Let δ_ε be chosen as in the proof of Lemma 2.2 and let $\varepsilon \in (0,1)$. If $\varphi(x,t) \in \mathrm{B}_{\delta_\varepsilon}(x^*)$, then (2.3) applied to $\varphi(x,t)$ (in the role of x) yields

$$\begin{aligned}\frac{\partial}{\partial t}\Big\|\varphi(x,t)-x^*\Big\| &= \big\langle v\big(\varphi(x,t)\big), \mathfrak{n}\big(\varphi(x,t)-x^*\big)\big\rangle \\ &\overset{(2.3)}{\leq} -(1-\varepsilon)\|\varphi(x,t)-x^*\| < 0.\end{aligned} \tag{3.5}$$

We now claim that

$$x \in \mathrm{B}_{\delta_\varepsilon}(x^*) \Rightarrow \varphi(x,t) \in \mathrm{B}_{\delta_\varepsilon}(x^*) \qquad \forall t \geq 0. \tag{3.6}$$

Indeed, if the contrary holds, then there exists $\tau > 0$ such that $\varphi(x,\tau) \in \partial\,\mathrm{B}_{\delta_\varepsilon}(x^*)$ and $\varphi(x,t) \in \mathrm{B}_{\delta_\varepsilon}(x^*)$ for all $t \in [0,\tau)$. But this leads to

$$\delta_\varepsilon = \|\varphi(x,\tau)-x^*\| = \|x-x^*\| + \int_0^\tau \frac{\partial}{\partial t}\Big|_{t=\theta}\|\varphi(x,t)-x^*\|\,d\theta < \delta_\varepsilon + \int_0^\tau 0\,d\theta,$$

which is a contradiction showing that (3.6) holds true, as claimed. It follows from (3.6) that $x \in \mathrm{B}_{\delta_\varepsilon}(x^*)$ implies (3.5) for all $t \geq 0$. Now let

$$R(x,t) := \frac{v\big(\varphi(x,t)\big) + \varphi(x,t) - x^*}{\|\varphi(x,t) - x^*\|}.$$

By the definition of δ_ε, $\varphi(x,t) \in \mathrm{B}_{\delta_\varepsilon}(x^*)$ implies

$$\|R(x,t)\| = \frac{\big\|\varphi(x,t) + v\big(\varphi(x,t)\big) - x^*\big\|}{\big\|\varphi(x,t) - x^*\|} < \varepsilon. \tag{3.7}$$

By (3.6), (3.7) thus holds true for all $t \geq 0$ when $x \in \mathrm{B}_{\delta_\varepsilon}(x^*)$, and then

$$\begin{aligned}\frac{\partial}{\partial t}\|\varphi(x,t) - x^*\| &= \big\langle v(\varphi(x,t)), \mathfrak{n}(\varphi(x,t) - x^*)\big\rangle \\ &= \big(-1 + \big\langle R(x,t), \mathfrak{n}(\varphi(x,t) - x^*)\big\rangle\big)\ \|\varphi(x,t) - x^*\| \\ &\overset{(3.7)}{\geq} -(1+\varepsilon)\|\varphi(x,t) - x^*\|.\end{aligned}$$

The combination of the last inequality with (3.5) establishes that if $x \in \mathrm{B}_{\delta_\varepsilon}(x^*)$, then

$$-(1+\varepsilon)\|\varphi(x,t) - x^*\| \leq \frac{\partial}{\partial t}\Big\|\varphi(x,t) - x^*\Big\| \leq -(1-\varepsilon)\|\varphi(x,t) - x^*\| \quad \forall t \geq 0. \tag{3.8}$$

Furthermore, for $x \in \mathrm{B}_{\delta_\varepsilon}(x^*)$ we have

$$\begin{aligned}\Big\|\frac{\partial}{\partial t}\mathfrak{n}\big(\varphi(x,t) - x^*\big)\Big\| &= \frac{\big\|v(\varphi(x,t)) - \frac{\partial}{\partial t}\|\varphi(x,t) - x^*\| \cdot \mathfrak{n}(\varphi(x,t) - x^*)\big\|}{\|\varphi(x,t) - x^*\|} \\ = \frac{\big\|v(\varphi(x,t)) + \varphi(x,t) - x^* - \big(\varphi(x,t) - x^* + \frac{\partial}{\partial t}\|\varphi(x,t) - x^*\| \cdot \mathfrak{n}(\varphi(x,t) - x^*)\big)\big\|}{\|\varphi(x,t) - x^*\|} &\qquad (3.9) \\ \leq \|R(x,t)\| + \frac{\big|\|\varphi(x,t) - x^*\| + \frac{\partial}{\partial t}\|\varphi(x,t) - x^*\|\big|}{\|\varphi(x,t) - x^*\|} \\ \overset{(3.7),(3.8)}{\leq} \varepsilon + \varepsilon. &\qquad (3.10)\end{aligned}$$

Equations (3.8), (3.10), and Lemma 3.1 therefore show that (3.2) and (3.3) hold with $\rho_\varepsilon = \delta_{\varepsilon/2}$ for all $\varepsilon \in (0,1)$. This settles the first claim of the lemma.

For the purposes of proving the second claim, let β and ρ be chosen as in (1.2), and note that if (1.2) holds, then by the same arguments as above the inequality (3.8) can be strengthened to

$$\begin{aligned}&-(1+\beta\|\varphi(x,t) - x^*\|^p)\|\varphi(x,t) - x^*\| \\ &\qquad \leq \frac{\partial}{\partial t}\Big\|\varphi(x,t) - x^*\Big\| \\ &\qquad\qquad \leq -(1-\beta\|\varphi(x,t) - x^*\|^p)\|\varphi(x,t) - x^*\|\end{aligned} \tag{3.11}$$

for $x \in \mathrm{B}_r(x^*)$ and $t \geq 0$, where $r = \min\{(\beta)^{-1/p}, \rho\}$ now ensures that if $x \in \mathrm{B}_r(x^*)$, then $\varphi(x,t) \in \mathrm{B}_r(x^*)$ for all $t \geq 0$. It follows from (3.11) that for $r_\varepsilon = \min\{(\varepsilon/\beta)^{1/p}, r\} = \min\{(\varepsilon/\beta)^{1/p}, \rho\}$, $x \in \mathrm{B}_{r_\varepsilon}(x^*)$, and $t \geq 0$, it is the case that

$$-(1+\varepsilon)\|\varphi(x,t) - x^*\| \leq \frac{\partial}{\partial t}\|\varphi(x,t) - x^*\| \leq -(1-\varepsilon)\|\varphi(x,t) - x^*\|.$$

Lemma 3.1 therefore shows that (3.2) holds with $\rho_\varepsilon = r_\varepsilon$. Furthermore, if (1.2) holds, then $x \in \mathrm{B}_{r_\varepsilon}(x^*)$ implies $\varphi(x,t) \in \mathrm{B}_{r_\varepsilon}(x^*) \subset \mathrm{B}_\rho(x^*)$ for all $t \geq 0$ and then

$$\|R(x,t)\| \leq \beta \|\varphi(x,t) - x^*\|^p.$$

Equations (3.11), (3.10), and (3.2) therefore show that (3.4) holds with $\gamma = 2\beta$. ∎

LEMMA 3.6. *If* (1.4) *converges exponentially to* x^**, then* (1.1) *converges uniformly Q-superlinearly to* x^**. Moreover, if* (1.4) *converges p-exponentially, then* (1.1) *converges uniformly at Q-convergence rate* $p+1$*.*

Proof. Let ρ_ε be as in Definition 3.3. Suppose that $\frac{\partial}{\partial t}\|\varphi(x,t_0) - x^*\| > -\|\varphi(x,t_0) - x^*\|(1-\varepsilon)$ for some $x \in \mathrm{B}_{\rho_\varepsilon}(x^*)$ and $t_0 > 0$. By continuity there exists $\delta > 0$ such that

$$\frac{\partial}{\partial t}\|\varphi(x,t) - x^*\| > -\|\varphi(x,t) - x^*\|(1-\varepsilon)$$

for all $t \in [t_0, t_0+\delta]$. By Lemma 3.1 we then have

$$\|\varphi(x,t) - x^*\| > \|\varphi(x,t_0) - x^*\| \, \mathrm{e}^{-(1-\varepsilon)t}$$

for $t \in [t_0, t_0+\delta]$, contradicting the upper bound in (3.2) when $\varphi(x,t_0)$ is used in place of x. This and a similar argument using the lower bound in (3.2) show that

$$-\|\varphi(x,t) - x^*\|(1+\varepsilon) \leq \frac{\partial}{\partial t}\|\varphi(x,t) - x^*\| \leq -\|\varphi(x,t) - x^*\|(1-\varepsilon)$$

for all $t \geq 0$ and $x \in \mathrm{B}_{\rho_\varepsilon}(x^*)$, and hence,

$$\begin{aligned} &\big\|P(x,t)\big(v(\varphi(x,t)) + \varphi(x,t) - x^*\big)\big\| \\ &\qquad = \left|\frac{\partial}{\partial t}\|\varphi(x,t) - x^*\| + \|\varphi(x,t) - x^*\|\right| \\ &\qquad \leq \varepsilon\|\varphi(x,t) - x^*\|. \end{aligned} \tag{3.12}$$

On the other hand, (3.1) and (3.3) show that for all $t \geq 0$ and $x \in \mathrm{B}_{\rho_\varepsilon}(x^*)$,

$$\begin{aligned} &\big\|(\mathrm{I} - P(x,t))\big(v(\varphi(x,t)) + \varphi(x,t) - x^*\big)\big\| \\ &\qquad = \Big\|\frac{\partial}{\partial t}\mathfrak{n}(\varphi(x,t) - x^*)\Big\| \cdot \|\varphi(x,t) - x^*\| \\ &\qquad \leq \varepsilon\|\varphi(x,t) - x^*\|. \end{aligned} \tag{3.13}$$

Inequalities (3.12) and (3.13) finally show that

$$\begin{aligned} &\frac{\big\|x + v(x) - x^*\big\|}{\|x - x^*\|} = \frac{\big\|v\big(\varphi(x,0)\big) + \varphi(x,0) - x^*\big\|}{\|\varphi(x,0) - x^*\|} \\ &\leq \frac{\big\|P(x,0)\big(v(\varphi(x,0)) + \varphi(x,0) - x^*\big)\big\|}{\|\varphi(x,0) - x^*\|} + \frac{\big\|\big(\mathrm{I} - P(x,0)\big)\big(v(\varphi(x,0)) + \varphi(x,0) - x^*\big)\big\|}{\|\varphi(x,0) - x^*\|} \\ &\leq 2\varepsilon \end{aligned} \tag{3.14}$$

for all $x \in \mathrm{B}_{\rho_\varepsilon}(x^*)$. Since this holds true for any $\varepsilon > 0$, we find that (1.3) holds, showing the first claim. If (1.4) converges p-exponentially, then the estimate (3.13) improves to

$$\Big\|\frac{\partial}{\partial t}\mathfrak{n}(\varphi(x,t) - x^*)\Big\| \cdot \|\varphi(x,t) - x^*\| \leq \gamma\|x - x^*\|^{p+1} \, \mathrm{e}^{-(1-\varepsilon)(p+1)t},$$

which leads to the inequality

$$\big\|(\mathrm{I}-P(x,0))\big(v(\varphi(x,0))+\varphi(x,0)-x^*\big)\big\| \le \gamma\|x-x^*\|^{p+1}. \tag{3.15}$$

Likewise, the estimate (3.12) improves, because there exists $\delta > 0$ such that for all $x \in \mathrm{B}_\delta(x^*)$ we have $x \in \mathrm{B}_{\rho_\varepsilon}(x^*)$ for some $\varepsilon \le \xi^{-p}\|x-x^*\|^p$. But then (3.2) shows that for $t \ge 0$,

$$\mathrm{e}^{-(1+\xi^{-p}\|x-x^*\|^p)t}\,\|x-x^*\| \le \|\varphi(x,t)-x^*\| \le \mathrm{e}^{-(1-\xi^{-p}\|x-x^*\|^p)t}\,\|x-x^*\|,$$

and by a similar argument as used above,

$$-\big(1+\xi^{-p}\|x-x^*\|^p\big)\|\varphi(x,t)-x^*\| \le \frac{\partial}{\partial t}\|\varphi(x,t)-x^*\| \le -\big(1-\xi^{-p}\|x-x^*\|^p\big)\|\varphi(x,t)-x^*\|$$

for all $t \ge 0$ and $x \in \mathrm{B}_\delta(x^*)$, so that

$$\left|\frac{\partial}{\partial t}\|\varphi(x,0)-x^*\| + \|\varphi(x,0)-x^*\|\right| \le \xi^{-p}\|x-x^*\|^{p+1}.$$

Using this in (3.12), we obtain

$$\big\|P(x,0)\big(v(\varphi(x,0))+\varphi(x,0)-x^*\big)\big\| \le \xi^{-p}\|x-x^*\|^{p+1}. \tag{3.16}$$

Finally, substituting (3.15) and (3.16) in (3.14), we obtain (1.2) with $\beta = \gamma+\xi^{-p}$ and $q = p$, and hence the second claim is true. □

4. Implications for the Newton process. The notion of exponential convergence shows that Q-convergence of the discrete process (1.1) corresponding to a vector field $v(x)$ is essentially due to two factors: the associated continuous flow $\varphi(x,t)$ converges exponentially fast to x^*, and the angular speed of $\varphi(x,t)$ relative to x^* decays to zero at a fast-enough rate. An important condition is that these rates must occur in neighborhoods of x^* that are not too small. Theorem 3.4 thus sets a geometric paradigm for constructing a vector field $v(x)$ such that the process (1.1) is attractive to x^* and converges fast. Let us now comment on the extent to which the Newton–Raphson approach satisfies this paradigm.

In Theorem 10.2.2 of [4], Ortega and Rheinboldt showed that if $f : D \to \mathbb{R}^n$, $x^* \in D \subset \mathbb{R}^n$, is such that $f(x^*) = 0$ and f' is p-Hölder continuous and nonsingular at x^*, then the Newton process converges to x^* at order at least $p+1$.

By linking the notion of uniform Q-convergence of order $p+1$ with the concept of p-exponential convergence, Theorem 3.4 can be used to derive an alternative proof that explains this phenomenon geometrically in the case where f is k times continuously differentiable for some $k > p \ge 1$: p-Hölder continuity of f' then implies that $p \in \mathbb{N}$ and $f''(x^*), \dots, f^{(p)}(x^*) = 0$ so that the Taylor developments of f, f', and f'' around x^* are of the form

$$f(x) = f'(x^*)[x-x^*] + f^{(p+1)}(x^*)[x-x^*,\dots,x-x^*] + o\left(\|x-x^*\|^{p+1}\right), \tag{4.1}$$

$$f'(x) = f'(x^*) + f^{(p+1)}(x^*)[x-x^*,\dots,x-x^*;\cdot] + o\left(\|x-x^*\|^{p}\right), \tag{4.2}$$

$$f''(x) = f^{(p+1)}(x^*)[x-x^*,\dots,x-x^*;\cdot,\cdot] + o\left(\|x-x^*\|^{p-1}\right). \tag{4.3}$$

Therefore,

$$\begin{aligned} v(x) &= -(f'(x))^{-1} f(x) \\ &= -\left(f'(x^*) + \mathscr{O}(\|x-x^*\|^p)\right)^{-1} \left(f'(x^*)(x-x^*) + \mathscr{O}(\|x-x^*\|^{p+1})\right) \\ &= -\left(f'(x^*)^{-1} + \mathscr{O}(\|x-x^*\|^p)\right) \left(f'(x^*)(x-x^*) + \mathscr{O}(\|x-x^*\|^{p+1})\right) \\ &= x^* - x + \mathscr{O}(\|x-x^*\|^{p+1}). \end{aligned} \tag{4.4}$$

Thus, multiplying the vector field $f(x)$ with $-(f'(x))^{-1}$ produces a new vector field $v(x)$ that asymptotically looks like the radial vector field $r(x) := x^* - x$ in the sense that

$$v(x) = r(x) + o(\|x-x^*\|). \tag{4.5}$$

This is exactly the condition needed to make the associated flux $\varphi(x,t)$ converge exponentially to x^*: (4.5) implies that for every $\varepsilon > 0$ there exists $\rho_\varepsilon > 0$ such that for all $x \in \mathrm{B}_{\rho_\varepsilon}(x^*)$,

$$\|v(x) - (x^* - x)\| \leq \varepsilon \|x - x^*\|. \tag{4.6}$$

Therefore,

$$\begin{aligned} \frac{\partial}{\partial t}\|\varphi(x,t)-x^*\| &= \|\varphi(x,t)-x^*\|^{-1} \left\langle \frac{\partial}{\partial t}\varphi(x,t), \varphi(x,t)-x^* \right\rangle \\ &= \langle v(\varphi(x,t)), \mathfrak{n}(\varphi(x,t)-x^*)\rangle \\ &\leq -\|\varphi(x,t)-x^*\| \cdot (1-\varepsilon), \end{aligned}$$

and likewise,

$$\frac{\partial}{\partial t}\|\varphi(x,t)-x^*\| \geq -\|\varphi(x,t)-x^*\| \cdot (1+\varepsilon).$$

Invoking Lemma 3.1, we find (3.2). Furthermore, by Lemma 3.2,

$$\left\| \frac{\partial}{\partial t}\mathfrak{n}(\varphi(x,t)-x^*) \right\| = \frac{\left\| \big(\mathrm{I} - P(\varphi(x,t))\big) v(\varphi(x,t)) \right\|}{\|\varphi(x,t)-x^*\|} \leq \varepsilon,$$

where the last inequality follows from $P(\varphi(x,t))(\varphi(x,t)-x^*) = \varphi(x,t)-x^*$ and (4.6). Therefore, (3.3) holds as required.

In order to furthermore achieve p-exponential convergence, the "rotational speed" $\frac{\partial}{\partial t}\mathfrak{n}(\varphi(x,t)-x^*)$ needs to decay uniformly and fast enough. But if $f'(x)$ is p-Hölder continuous, then

$$\begin{aligned} \frac{\partial}{\partial t}\Big|_{t=0} \mathfrak{n}(\varphi(x,t)-x^*) &\overset{(3.1)}{=} \frac{(\mathrm{I}-P(x,0))\, v(x)}{\|x-x^*\|} \\ &\overset{(4.4)}{=} (\mathrm{I}-P(x,0)) \left(\frac{x-x^* + \mathscr{O}(\|x-x^*\|^{p+1})}{\|x-x^*\|} \right) \\ &= \mathscr{O}(\|x-x^*\|^p), \end{aligned}$$

where we used $(\mathrm{I}-P(x,0))(x-x^*) = 0$. This shows that the rotational component dies out at the required rate (3.4).

5. Q-convergence rates for rescaled vector fields. The question naturally arises as to whether a rescaled vector field $\lambda(x)v(x)$ for some scalar function $\lambda : D \to \mathbb{R}_+$ has a faster Q-convergence rate than $v(x)$. Such rescaling is routinely used in unconstrained optimization, where search-direction-based descent methods are combined with line-searches. It turns out that in the general case the Q-convergence rate cannot be arbitrarily increased by such a rescaling, that is, there exist vector fields $v \in C^1(D, \mathbb{R}^n)$ for which there exists a bound $\bar{q} > 0$ such that for all positive scalar functions $\lambda \in C^1(D, \mathbb{R}_+)$ the vector field $\lambda(x)v(x)$ has a Q-convergence rate no faster than $\bar{q}$. Indeed, the vector field of Example 5.1 has this property, as shown in Theorem 5.3.

We believe that the existence of a finite $\bar{q}$ is in fact typical in the sense that arbitrarily fast convergence rates can be achieved only for a thin set of vector fields, where thin has to be appropriately defined. For the moment, rather than developing such a theory here, we content ourselves with answering the more modest question regarding the existence of vector fields with finite $\bar{q}$.

Example 5.1. Let $g(t) := \mathrm{e}^{-t}\,\mathrm{e}^{-1/t}$. Then the vector field

$$v(x) := \begin{cases} 0 & \text{if } x = 0, \\ -x & \text{if } x \in \mathrm{S}^1 = \partial\,\mathrm{B}_1(0), \\ -\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + g'\left(-\ln\|x\|\right)\begin{bmatrix} -x_2 \\ x_1 \end{bmatrix} & \text{if } x \in \mathrm{B}_1(0) \setminus \{0\} \end{cases}$$

is continuously differentiable on the closed unit ball $\overline{\mathrm{B}_1(0)}$, and furthermore we have $J(0) = -\,\mathrm{I}$, that is, $v(x)$ satisfies the conditions of Lemma 2.1 for superlinear convergence of the process (1.1) to the stationary point $x^* = 0$.

It is easily checked by taking the partial derivative with respect to t that for $y \in \mathrm{S}^1$, the flux line through y defined by the vector field $v(x)$ defined in Example 5.1 is given by

$$\varphi(y,t) = \mathrm{e}^{-t}\begin{bmatrix} y_1 & -y_2 \\ y_2 & y_1 \end{bmatrix}\begin{bmatrix} \cos g(t) \\ \sin g(t) \end{bmatrix}. \tag{5.1}$$

We also observe that $\varphi(U_\theta y, t) = U_\theta \varphi(y,t)$ holds for all rotations

$$U_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix};$$

see Figure 5.1. Since $\lim_{t\to\infty}\varphi(y,t) = 0$ for all $y \in \mathrm{S}^1$, this rotational invariance implies that for all $x \in \mathrm{B}_1(0)$ there exists $y^{[x]} \in \mathrm{S}^1$ such that

$$\varphi(x,t) = \mathrm{e}^{-t}\,\|x\|\begin{bmatrix} y_1^{[x]} & -y_2^{[x]} \\ y_2^{[x]} & y_1^{[x]} \end{bmatrix}\begin{bmatrix} \cos g(t - \ln\|x\|) \\ \sin g(t - \ln\|x\|) \end{bmatrix}.$$

To find $y^{[x]}$, it suffices to integrate the flow in reverse direction starting from x until the flux line crosses S^1. This crossing happens in finite time because of (5.1) and the flow-conservation property (1.5). It is also easy to check that

$$\left\| \frac{\partial}{\partial t}\mathfrak{n}(\varphi(x,t)) \right\| = \left| g'\left(t - \ln\|x\|\right) \right|, \tag{5.2}$$

$$\|\varphi(x,t)\| = \mathrm{e}^{-t}\,\|x\| \tag{5.3}$$

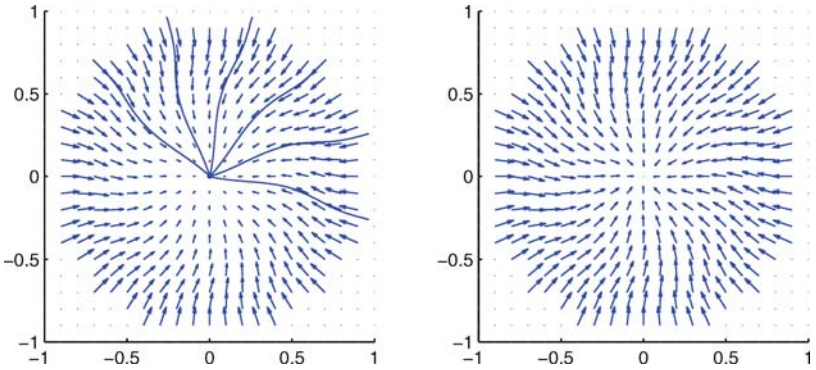


FIG. 5.1. *The figure on the left shows the vector field of Example* 5.1 *and a few of its flux lines, while the figure on the right shows a rescaling of the same vector field, here* $w(x) = v(x)/\sqrt{\|x\|}$.

for all $t \geq 0$ and $x \in \mathrm{B}_1(0)$.

LEMMA 5.2. *For the vector field* $v(x)$ *constructed in Example* 5.1 *we have*

$$\bar{q} := \sup\{q : \text{the process (1.1) defined by } v(x) \text{ is unif. Q-conv. with rate } q\} < +\infty.$$

Proof. It suffices to show that the process (1.1) is uniformly Q-convergent at rate 2 but not 3. By Theorem 3.4 this is equivalent to showing that the continuous process (3) converges 1-exponentially to $x^* = 0$ but not 2-exponentially. Note that

$$\|\varphi(x,t)\| = \mathrm{e}^{-t}\ \|x\|.$$

To satisfy (3.2), ρ_ε can thus be chosen arbitrarily in $(0,1)$, for example, $\rho_\varepsilon = \mathrm{e}^{-1}$, so that $\rho_\varepsilon \geq \varepsilon = 1 \cdot \varepsilon^{1/1}$ for all $\varepsilon \leq \mathrm{e}^{-1}$, as required for the claim of 1-exponential convergence. Furthermore, we have

$$\begin{aligned} \left\| \frac{\partial}{\partial t} \mathfrak{n}\left(\varphi(x,t) - x^*\right) \right\| &= |g'\left(t - \ln\|x\|\right)\| \\ &= \|x\| \cdot \mathrm{e}^{\frac{1}{\ln\|x\| - t}} \cdot \left| 1 - \frac{1}{(t - \ln\|x\|)^2} \right| \cdot \mathrm{e}^{-t} \end{aligned} \tag{5.4}$$

for all $t \geq 0$, and since $x \in \mathrm{B}_{\mathrm{e}^{-1}}(0)$ implies

$$\mathrm{e}^{\frac{1}{\ln\|x\| - t}} \cdot \left| 1 - \frac{1}{(t - \ln\|x\|)^2} \right| \leq 1,$$

we find that (3.4) holds with $p = 1$ and $\gamma = 1$. This establishes the claimed 1-exponential convergence. On the other hand, if 2-exponential convergence were to hold, then

$$\left\| \frac{\partial}{\partial t} \mathfrak{n}\left(\varphi(x,t)\right) \right\| \leq \gamma \cdot \|x\|^2 \cdot \mathrm{e}^{-2(1-\varepsilon)t}$$

would have to be true for some fixed γ and for all $t \geq 0$, $x \in \mathrm{B}_{\xi\varepsilon^{1/p}}(0)$, and $\varepsilon \in (0,1)$ for some appropriately chosen $\xi > 0$. By virtue of (5.4), we thus need

$$\mathrm{e}^{\frac{1}{\ln\|x\| - t}} \cdot \left| 1 - \frac{1}{(t - \ln\|x\|)^2} \right| \leq \gamma \cdot \|x\| \cdot \mathrm{e}^{-t + 2\varepsilon t}$$

to hold for all $t \geq 0$ and $x \in \mathrm{B}_{\rho_\varepsilon}(0)$. But since the left-hand side tends to 1 when $t \to \infty$ while the right-hand side converges to zero when $\varepsilon < 1/2$, this cannot be achieved for any fixed γ. ∎

We are now ready to show that a rescaling of the vector field $v(x)$ from Example 5.1 cannot speed up the Q-convergence rate of the process (1.1).

THEOREM 5.3. *Let $v(x)$ and $\bar{q}$ be as in Lemma 5.2, let $\lambda \in C^1(\mathbb{R}^2, \mathbb{R}_+)$ be a positive scalar function and consider the vector field $w(x) = \lambda(x)v(x)$. Then*

$$\sup\{q : \textit{the process } (1.1) \textit{ defined by } w(x) \textit{ is unif. Q-conv. with rate } q\} \leq \bar{q}.$$

Proof. Let $\bar{p} = \bar{q} - 1$. It suffices to show that for $p > \bar{p}$ the process (1.1) $w(x)$ is not Q-convergent with rate $p+1$. If the contrary holds for some choice of $\lambda(x)$, then the flux $\psi(x,t)$ associated with $w(x)$ converges p-exponentially to the origin, that is, there exist constants $\xi > 0$, $\gamma_p > 0$ such that for all $\varepsilon > 0$ small enough there exists $\rho_\varepsilon > \xi\varepsilon^{1/p}$ with the property that for all $x \in \mathrm{B}_{\rho_\varepsilon}(0)$ and $t \geq 0$,

$$\mathrm{e}^{-(1+\varepsilon)t}\,\|x\| \leq \|\psi(x,t)\| \leq \mathrm{e}^{-(1-\varepsilon)t}\,\|x\|, \tag{5.5}$$

$$\left\| \frac{\partial}{\partial t}\mathfrak{n}(\psi(x,t)) \right\| \leq \gamma_p \|x\|^p\,\mathrm{e}^{-(1-\varepsilon)pt}\,. \tag{5.6}$$

On the other hand, $\varphi(x,t)$ does not converge p-exponentially, and since (5.3) holds for all $x \in \mathrm{B}_1(0)$, this must be because for $0 < \varepsilon \ll 1$ there exists $z_\varepsilon \in \mathrm{B}_{\rho_\varepsilon}(0)$ and $T \geq 0$ such that

$$\frac{\gamma_p}{1-\varepsilon}\|z_\varepsilon\|^p\,\mathrm{e}^{-(1-\varepsilon)pT} < \left|g'\left(T - \ln\|z_\varepsilon\|\right)\right|.$$

Replacing z_ε by $\varphi(z_\varepsilon, T)$, it is easy to see that we may assume without loss of generality that $T = 0$, and then by continuity there exist constants $\delta, \tau > 0$ such that

$$\frac{\gamma_p}{1-\varepsilon}\|z_\varepsilon\|^p\,\mathrm{e}^{-(1-\varepsilon)pt} < \left|g'\left(t - \ln\|z_\varepsilon\|\right)\right|(1-\delta) \tag{5.7}$$

for all $t \in [0, \tau)$. Now let $s(t)$ be defined by the ODE

$$\frac{d}{dt}s(t) = \lambda\big(\varphi(z_\varepsilon, t)\big), \qquad s(0) = 0.$$

Since $\lambda > 0$, $s(t)$ is a monotone increasing reparameterization of t constructed so that $\psi(z_\varepsilon, t) = \varphi\big(z_\varepsilon, s(t)\big)$. Using the chain rule, we find

$$\begin{aligned}\left\| \frac{\partial}{\partial t}\mathfrak{n}(\psi(z_\varepsilon,t)) \right\| &= \left|\frac{d}{dt}s(t)\right| \cdot \left\| \frac{\partial}{\partial s}\mathfrak{n}(\varphi(z_\varepsilon,s)) \right\| \\ &\overset{(5.2)}{=} \frac{d}{dt}s(t) \cdot \left|g'\left(s(t) - \ln\|z_\varepsilon\|\right)\right|,\end{aligned}$$

and using (5.6)) and (5.7) we obtain that

$$\begin{aligned}\frac{d}{dt}s(t) \cdot \frac{\gamma_p}{1-\varepsilon}\,\|z_\varepsilon\|^p\,\mathrm{e}^{-(1-\varepsilon)p\,s(t)} &< \frac{d}{dt}s(t) \cdot \left|g'\left(s(t) - \ln\|z_\varepsilon\|\right)\right|\,(1-\delta) \\ &\leq \gamma_p\,\|z_\varepsilon\|^p\,\mathrm{e}^{-(1-\varepsilon)pt}\,(1-\delta)\end{aligned}$$

holds for every $t \in [0, s^{-1}(\tau))$, that is,

$$\frac{d}{dt}\left(s(t)-t\right) < \frac{\gamma_p \left\|z_\varepsilon\right\|^p \, \mathrm{e}^{-(1-\varepsilon)pt}\,(1-\delta)}{\frac{\gamma_p}{1-\varepsilon}\left\|z_\varepsilon\right\|^p \, \mathrm{e}^{-(1-\varepsilon)p\,s(t)}} - 1 = (1-\varepsilon)\,(1-\delta)\,\mathrm{e}^{(1-\varepsilon)p\,(s(t)-t)} - 1.$$

Note that the right-hand side is monotone increasing in $s(t)-t$. Therefore, we can apply Lemma 3.1 to find that

$$s(t) - t < -\frac{\ln\left[(1-\varepsilon)(1-\delta) + (\varepsilon+\delta-\varepsilon\delta)\,\mathrm{e}^{(1-\varepsilon)pt}\right]}{(1-\varepsilon)p}$$

holds for small positive t. Since

$$\ln\left[(1-\varepsilon)(1-\delta) + (\varepsilon+\delta-\varepsilon\delta)\,\mathrm{e}^{(1-\varepsilon)pt}\right] = (\varepsilon+\delta-\varepsilon\delta)(1-\varepsilon)pt + \mathscr{O}(t^2),$$

it follows that

$$s(t) - t < -\varepsilon t \tag{5.8}$$

for small positive t. On the other hand, (5.3) and (5.5) imply

$$\mathrm{e}^{-s(t)}\left\|z_\varepsilon\right\| = \left\|\varphi(z_\varepsilon, s(t))\right\| = \left\|\psi(z_\varepsilon, t)\right\| \le \mathrm{e}^{-(1-\varepsilon)t}\left\|z_\varepsilon\right\|,$$

and hence, $-\varepsilon t \le s(t) - t$ for small positive t. Since this contradicts (5.8), it follows that our claim is true. ∎

Acknowledgments. The authors would like to thank the three referees and the associate editor for extremely helpful feedback and valuable suggestions for improving the paper.

REFERENCES

[1] R. Courant, *Variational methods for the solution of problems of equilibrium and vibrations*, Bull. Amer. Math. Soc., 49, (1943), pp. 1–23.
[2] U. Helmke and J. B. Moore, *Optimization and Dynamical Systems*, Springer-Verlag, London, 1994.
[3] L.V. Kantorovich, *Functional analysis and applied mathematics,* Uspehi Mat. Nauk, 3 (1948), pp. 89–185, (in Russian).
[4] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.
[5] F. Potra, *On Q-order and R-order of convergence*, J. Optim. Theory Appl., 63 (1989), pp. 415–431.
[6] J. Szarski, *Differential Inequalities*, 2nd rev. ed., Monografie Matematyczne 43, PWN–Polish Scientific Publishers, Warszawa, 1967.

SIAM J. OPTIM.
Vol. 18, No. 1, pp. 65–86

AN INTERIOR-POINT TRUST-REGION ALGORITHM FOR GENERAL SYMMETRIC CONE PROGRAMMING*

YE LU† AND YA-XIANG YUAN‡

Abstract. An interior-point trust-region algorithm is proposed for minimizing a general (nonconvex) quadratic objective function in the intersection of a symmetric cone and an affine subspace. The algorithm uses a trust-region model to ensure descent on a suitable merit function. Global first-order and second-order convergence results are proved. Numerical results are presented.

Key words. interior-point algorithm, trust-region subproblem, symmetric cone

AMS subject classifications. 90C30, 90C51

DOI. 10.1137/040611756

1. Introduction. In the last two decades, interior-point algorithms for convex programming have been developed quite well in both theory and practice. However, research on interior-point algorithms for nonconvex programming is still very active, as nonconvex problems are considerably more difficult. We mention here some of the recent works. For semidefinite relaxations, see Zhang [33] and Ye and Zhang [31]; for line search based algorithms, see Absil and Tits [1] and Bakry et al. [3], Forsgren and Gill [12], Gay, Overton, and Wright [13], Tits et al. [25], Vanderbei and Shanno [27], and Wächter [28]. By contrast, for trust-region-type interior-point algorithms, Ye [29] developed an affine scaling algorithm for indefinite quadratic programming by solving sequential trust-region subproblems. Global first-order and second-order convergence results were proved, and later enhanced by Sun [24] for the convex case. The idea of affine scaling can be traced back to Dikin [8]. An affine-scaling potential-reduction interior-point trust-region algorithm was developed for the indefinite quadratic programming in Ye [30, section 9]. Recently, in Faybusovich and Lu [11], Ye's algorithm has been extended to the minimization of a quadratic function in the intersection of a symmetric cone and an affine subspace. In this paper, we call such a problem symmetric cone programming and develop an affine-scaling primal barrier interior-point trust-region algorithm to solve it. Since the class of symmetric cones contains the positive orthant in R^n, the second-order cone, and the cone of positive semidefinite symmetric matrices, our approach solves a large class of optimization problems. In the trust-region literature, we refer the reader to Conn, Gould, and Toint [6, section 13] for a primal barrier algorithm and Conn et al. [7] for a primal-dual algorithm. Under the theoretical framework of their work, we bring the properties of ϑ-*normal barrier* and symmetric cone into our analysis. By doing so, we show that the primal barrier algorithm developed in Conn, Gould, and Toint [6] can be extended to solve symmetric cone programming. Although our algorithm still provides the mechanism to declare the iteration unsuccessful if feasibility is not achieved, it does not contain an explicit

*Received by the editors July 16, 2004; accepted for publication (in revised form) August 24, 2006; published electronically February 2, 2007.
http://www.siam.org/journals/siopt/18-1/61175.html

†Operations Research Center, Massachusetts Institute of Technology, 77 Mass Ave., Bldg. E40-130, Cambridge, MA 02139 (yelu@mit.edu). This work was done while the author was a graduate student at the University of Notre Dame Mathematics Department.

‡LSEC, ICMSEC, AMSS, Chinese Academy of Sciences, Beijing 100080, China (yyx@lsec.cc.ac.cn). This author's work was partially supported by NSFC grant 10231060.

constraint on the step calculation (see constraint (13.2.1) in Conn, Gould, and Toint [6, p. 499] and the constraint (13.3.1) of the algorithm in Conn, Gould, and Toint [6, p. 505]). This makes our algorithm theoretically somewhat simpler, although its practical merit remains to be investigated. Moreover, we establish inequality (4.22) in section 4 of this paper to explicitly estimate the convergence of the algorithm. This is quite remarkable, since our analysis does not require any convexity assumptions.

This paper is organized as follows. In section 2, we present some concepts and results of the symmetric cone and *ϑ-normal barrier* in the theory of interior-point methods. In section 3, we formulate the first-order and second-order optimality conditions for our optimization problem. In section 4, we present a convergence analysis for our interior-point trust-region algorithm. The techniques of proofs in Lemmas 4.3–4.6 essentially follow from Conn, Gould, and Toint [6, section 13], together with the applications of the properties of the *ϑ-normal barrier*. In section 5, our algorithm is used to solve the large-scale trust-region subproblem. In section 6, we apply our algorithm to a class of quadratic programs and discuss some further implementation issues. Concluding remarks and recommendations are presented in section 7.

2. Symmetric cone and ϑ-normal barrier. In this section we introduce some concepts and relevant results which will be used in the following sections.

Nesterov and Nemirovskii [18] developed the concept of *ϑ-normal barrier*, which has become one of the most important tools for the analysis of interior-point methods. It is also an essential tool in our analysis. We assume that K is a convex cone in a finite-dimensional real vector space E. Let K° be the interior of K. The definition of *ϑ-normal barrier* is given as follows.

DEFINITION 2.1. *Let $F : K^\circ \to R$ be a C^3-smooth strictly convex function such that F is a barrier for K (i.e., $F(x) \to \infty$ as $x \in K^\circ$ approaches the boundary of K), and there exists $\vartheta \geq 1$ such that for each $t > 0$,*

$$F(tx) = F(x) - \vartheta \ln(t) \tag{2.1}$$

and

$$| F^{'''}(x)[h,h,h] | \leq 2 \langle F^{''}(x)h, h\rangle^{3/2} \tag{2.2}$$

for all $x \in K^\circ$ and for all $h \in E$. Then F is called a ϑ-normal barrier for K and ϑ is called barrier parameter of F.

In principle, every convex cone admits a *ϑ-normal barrier* (see Nesterov and Nemirovskii [18, section 4]). But, in this paper we consider only a special kind of convex cone, called a symmetric cone. As a regular convex cone K in a finite-dimensional real vector space E endowed with an inner product $\langle\ \rangle$, the dual of K is defined as

$$K^* = \{y \in E \mid \langle x, y\rangle \geq 0\ \forall x \in E\}.$$

We define $Aut(K)$ to be the set of automorphisms of the convex cone K, that is, $AK = K$ for any $A \in Aut(K)$. The following is the definition of symmetric cone.

DEFINITION 2.2. *A convex cone K is called homogeneous if $Aut(K)$ is transitive on K°; that is, given any pair of points $x, s \in K^\circ$ there exists $A \in Aut(K)$ such that $Ax = s$. The cone K is said to be self-dual if there is an inner product such that $K^* = K$. K is said to be symmetric if it is homogeneous and self-dual.*

The following important cones are special symmetric cones.

The positive orthant. The simplest symmetric cone is the positive orthant

$$R^n_{++} = \{x \mid x > 0, x \in R^n\} = R_{++} \oplus \cdots \oplus R_{++}, \tag{2.3}$$

which is the direct sum of n copies of R_{++}. $F(x) = -\sum_{i=1}^{n} \ln x_i$ is a *ϑ-normal barrier* for R^n_{++} with $\vartheta = n$.

The second-order cone. This is the cone defined by

$$SOC := \left\{ x \in R^n : \sum_{i=1}^{n-1} x_i^2 \le x_n^2 \text{ and } x_n \ge 0 \right\}. \tag{2.4}$$

The function $F(x) = -\ln(x_n^2 - \sum_{i=1}^{n-1} x_i^2)$ is a *ϑ-normal barrier* for the second-order cone SOC with $\vartheta = 2$.

The cone of positive semidefinite matrices. This is the cone of all positive semidefinite matrices

$$S_+^{n\times n} = \{X \mid X \in R^{n\times n}, \quad X \text{ positive semidefinite}\}. \tag{2.5}$$

$F(X) = -\ln\det(X)$ is a *ϑ-normal barrier* for $S_+^{n\times n}$ with $\vartheta = n$.

Let $F''(x)$ denote the Hessian of *ϑ-normal barrier* $F(x)$. The strictly convex assumption of $F(x)$ implies that $F''(x)$ is positive definite for every $x \in K^\circ$. Thus, $\|v\|_x = \langle v, F''(x)v\rangle^{\frac{1}{2}}$ is a norm on E induced by $F''(x)$. Let $B_x(y, r)$ denote the open ball of radius r centered at y, where the radius is measured with respect to $\| \ \|_x$. This ball is called the Dikin ball. The following lemmas are very crucial for the analysis of our algorithm in the next sections.

LEMMA 2.1. *Assume $F(x)$ is a ϑ-normal barrier for K; then for all $x \in K^\circ$ we have $B_x(x, 1) \subseteq K^\circ$.*

LEMMA 2.2. *Assume $F(x)$ is a ϑ-normal barrier for K, $x \in K^\circ$, and $y \in B_x(x, 1)$; then*

$$\left| F(y) - F(x) - \langle F'(x), y - x\rangle - \frac{\langle y - x, F''(x)(y - x)\rangle}{2} \right| \le \frac{\|y - x\|_x^3}{3(1 - \|y - x\|_x)}. \tag{2.6}$$

LEMMA 2.3. *Let F be a ϑ-normal barrier for K; then*

$$F''(x)^{-1}F'(x) = -x, \tag{2.7}$$

$$\langle -F'(x), x\rangle = \vartheta. \tag{2.8}$$

LEMMA 2.4. *If K is a symmetric cone and F is a ϑ-normal barrier for K, then $F''(x)$ is a linear automorphism of K for each $x \in K^\circ$.*

The proofs of the above lemmas can be found in Chapter 2 of Renegar [20].

Lemma 2.1 tells us the ball of radius 1 measured by $\|.\|_x$ is always contained in K°. Lemma 2.2 shows that at least locally, the quadratic approximation is very good for the *ϑ-normal barrier* F. Lemma 2.3 plays an important role in our proof of Lemma 4.1 in section 4. Lemma 2.4 is a special property of the symmetric cone, which is one of the reasons why in this paper we focus on the symmetric cones instead of general cones.

3. Optimality conditions. In this section, we formulate the first-order and second-order optimality conditions of our optimization problem:

We consider the following optimization problem:

$$\min \quad q(x) = \frac{1}{2}\langle x, Qx\rangle + \langle c, x\rangle \tag{3.1}$$

$$\text{subject to} \quad Ax = b, \tag{3.2}$$

$$x \in K. \tag{3.3}$$

Here $Q : E \mapsto E$ is a symmetric linear operator, $c \in E$. $A : E \mapsto R^m$ is a linear operator and $b \in R^m$. K is a symmetric cone. We assume that our feasible set $\digamma_p = \{x \in E | Ax = b, x \in K\}$ is bounded and has relative interior. The following theorem is the first-order optimality condition for our optimization problem. For a proof, see, e.g., Bonnans and Shapiro [5] or Faybusovich and Lu [11].

THEOREM 3.1 (first-order optimality condition). *If x^* is a locally minimal solution of* (3.1)–(3.3), *then there exists $s \in K^*(= K)$ such that $Qx^* + c - s \in R(A^*)$ and $\langle x^*, s\rangle = 0$; here $R(A^*)$ is the range of A^* and $A^* : R^m \mapsto E$ is the adjoint of A.*

Assume x is a point in our feasible set $\digamma_p = \{x \in E | Ax = b, x \in K\}$; there must be a unique face $\digamma_x$ of $\digamma_p$ such that x is a relative interior point of $\digamma_x$. We denote $Aff(\digamma_x)$ to be the affine space generated by $\digamma_x$ and V_x to be the vector space such that $Aff(\digamma_x) = V_x + x$. Now we are ready to formulate the second-order optimality condition.

THEOREM 3.2 (second-order optimality condition). *If x^* is a locally minimal solution of* (3.1)–(3.3), *$\digamma_{x^*}$ is the unique face of the feasible set $\digamma_p$ such that x^* is one of its relative interior points, and $V_{x^*} = Aff(\digamma_{x^*}) - x^*$, then Q is positive semidefinite over V_{x^*}.*

Proof. For all $d \in V_{x^*}$, because x^* is a relative interior of $\digamma_{x^*}$, we know $x^* + td \in \digamma_{x^*}$, provided that $|t|$ is sufficiently small. Hence, there exists a $\epsilon > 0$ such that

$$q(x^* + td) - q(x^*) = t\langle Qx^* + c, d\rangle + \frac{t^2}{2}\langle d, Qd\rangle \geq 0 \tag{3.4}$$

as long as $|t| \leq \epsilon$, due to the fact that x^* is a local minimim. The above inequality implies that

$$\langle Qx^* + c, d\rangle = 0, \qquad \langle d, Qd\rangle \geq 0. \tag{3.5}$$

This completes our proof. ∎

If $x \in K^\circ$, it is obvious that $V_x = \{x \in E | Ax = 0\}$. If $x \in \partial K$, the matter becomes much more complicated. But fortunately, we can get some very helpful results in the case of symmetric cones.

If $K = R^n_{++}$ and $x' \in \partial K$, it can be shown that $V_{x'} = \{x \in R^n | Ax = 0, x_j = 0, j \in I\}$, where $I = \{j | x'_j = 0\}$. We know that $F(x) = -\sum_{i=1}^n \ln x_i$ is a *ϑ-normal barrier* for R^n_{++}. Therefore, $F''(x')^{-\frac{1}{2}} = \mathrm{diag}\{x'_1, x'_2, \ldots, x'_n\}$, and consequently $V_{x'} = \{F''(x')^{-\frac{1}{2}} x | AF''(x')^{-\frac{1}{2}} x = 0, , x \in R^n\}$.

If $K = S^{n\times n}_+$, we know $F(X) = -\ln\det(X)$ is a *ϑ-normal barrier* for $S^{n\times n}_+$. Now let $A' \in \partial K$, and $\mathrm{rank}(A') = r < n$. Then $F''(A')^{-\frac{1}{2}} X = A'^{\frac{1}{2}} X A'^{\frac{1}{2}}$. We set $V = \{A'^{\frac{1}{2}} X A'^{\frac{1}{2}} | AA'^{\frac{1}{2}} X A'^{\frac{1}{2}} = 0, X \in S^{n\times n}\}$; just as with the positive orthant case, it holds that $V_{A'} = V$. The following theorem tells us that this property actually holds for all symmetric cones. We will prove this theorem in the appendix.

THEOREM 3.3. *Assume K is a symmetric cone in a finite-dimensional real Euclidean space E and $F(x)$ is the ϑ-normal barrier for K. If $x^* \in K$, then $V_{x^*} = \{F''(x^*)^{-\frac{1}{2}} x | AF''(x^*)^{-\frac{1}{2}} x = 0, x \in E\}$.*

We want to mention that $F''(x^*)^{-\frac{1}{2}}$ is well defined on the boundary of the cone, since it is the quadratic representation of x^* in Jordan algebra. This theorem immediately implies the following corollary, which is extremely important to prove that any limit point of the iterate generated by our algorithm satisfies the second-order optimality condition.

COROLLARY 3.1. *Assume K is a symmetric cone in our optimization problem (3.1)–(3.3); then Q is positive semidefinite on V_{x^*} if and only if the linear operator $F''(x^*)^{-\frac{1}{2}}QF''(x^*)^{-\frac{1}{2}}$ is positive semidefinite on $\{x|AF''(x^*)^{-\frac{1}{2}}x = 0, x \in E\}$.*

4. Interior-point trust-region algorithm. In this section, we present our interior-point trust-region algorithm for solving (3.1)–(3.3). Global first-order and second-order convergence results are proved.

We assume $F(x)$ is the *ϑ-normal barrier* for the symmetric cone K and define the merit function as

$$f_{\eta_k}(x) = q(x) + \frac{1}{\eta_k}F(x). \tag{4.1}$$

In the inner iterations, $f_{\eta_k}(x)$ is decreased for a fixed η_k, while η_k is increased to positive infinity in outer iterations. From Lemma 2.1, for any $x_{k,j} \in K^\circ$ and $d \in E$, we have that $x_{k,j} + d \in K^\circ$, provided that $\|F''(x_{k,j})^{\frac{1}{2}}d\| \le \alpha_{k,j} < 1$. It follows from Lemma 2.2 that

$$\begin{aligned} F(x_{k,j}+d) - F(x_{k,j}) &\le \langle F'(x_{k,j}), d\rangle + \frac{\langle d, F''(x_{k,j})d\rangle}{2} + \frac{\|d\|^3_{x_{k,j}}}{3(1-\|d\|_{x_{k,j}})} \\ &\le \langle F'(x_{k,j}), d\rangle + \frac{\langle d, F''(x_{k,j})d\rangle}{2} + \frac{\alpha^3_{k,j}}{3(1-\alpha_{k,j})}. \end{aligned} \tag{4.2}$$

Therefore, we get

$$\begin{aligned} f_{\eta_k}(x_{k,j}+d) - f_{\eta_k}(x_{k,j}) \le{}& \frac{\langle d, (Q + \frac{1}{\eta_k}F''(x_{k,j}))d\rangle}{2} \\ &+ \left\langle Qx_{k,j} + c + \frac{1}{\eta_k}F'(x_{k,j}), d\right\rangle + \frac{\alpha^3_{k,j}}{3(1-\alpha_{k,j})\eta_k}. \end{aligned} \tag{4.3}$$

From the above relation, it is obvious that in order to decrease $f_{\eta_k}(x)$, we can try to minimize its upper bound given in the right-hand side of the above inequality. This leads to the following subproblem:

$$\min \frac{1}{2}\left\langle d, \left(Q + \frac{1}{\eta_k}F''(x_{k,j})\right)d\right\rangle + \left\langle Qx_{k,j} + c + \frac{1}{\eta_k}F'(x_{k,j}), d\right\rangle = m_{k,j}(d) \tag{4.4}$$

$$\text{subject to} \quad Ad = 0, \tag{4.5}$$

$$\|F''(x_{k,j})^{\frac{1}{2}}d\|^2 \le \alpha^2_{k,j}. \tag{4.6}$$

Define

$$Q_{k,j} = F''(x_{k,j})^{-\frac{1}{2}}QF''(x_{k,j})^{-\frac{1}{2}} + \frac{1}{\eta_k}I, \tag{4.7}$$

$$c_{k,j} = F''(x_{k,j})^{-\frac{1}{2}}\left(Qx_{k,j} + c + \frac{1}{\eta_k}F'(x_{k,j})\right), \tag{4.8}$$

$$A_{k,j} = AF''(x_{k,j})^{-\frac{1}{2}}, \tag{4.9}$$

and using the transformation

$$d' = F''(x_{k,j})^{\frac{1}{2}}d, \tag{4.10}$$

equations (4.4)–(4.6) can be rewritten as

$$\min q'_{k,j}(d') = \frac{1}{2}\langle d', Q_{k,j}d'\rangle + \langle c_{k,j}, d'\rangle \tag{4.11}$$

$$\text{subject to} \qquad A_{k,j}d' = 0, \tag{4.12}$$

$$\|d'\|^2 \le \alpha_{k,j}^2. \tag{4.13}$$

Instead of solving (4.11)–(4.13) exactly, we need only compute an approximate solution $d'_{k,j}$ satisfying the following two inequalities:

$$q'_{k,j}(d'_{k,j}) \le -\theta\|p_{k,j}\| \min\left\{\frac{\|p_{k,j}\|}{\beta_{k,j}}, \alpha_{k,j}\right\} \tag{4.14}$$

and

$$q'_{k,j}(d'_{k,j}) \le \theta\lambda_{k,j}\min\{\lambda_{k,j}^2, \alpha_{k,j}^2\}, \tag{4.15}$$

where $\theta \in (0, \frac{1}{2})$, $\beta_{k,j} = 1 + \|Q_{k,j}\|$, $p_{k,j}$ is the projection of $c_{k,j}$ onto the null space of $A_{k,j}$, and $\lambda_{k,j}$ is the least eigenvalue of $(N_{k,j})^* Q_{k,j} N_{k,j}$, $N_{k,j}$ being an orthonormal basis spanning the null space of $A_{k,j}$. We can see that inequality (4.15) makes sense only when $\lambda_{k,j} < 0$. The two conditions (4.14) and (4.15) are common in trust-region methods. Inequality (4.14) can be obtained at the Cauchy point and inequality (4.15) can be obtained when the negative curvature is exploited. Projected conjugate gradient/Lanczos-like methods are able to produce such a step at a reasonable cost (see Gould et al. [14]).

Once $d'_{k,j}$ is computed, we obtain the trial step

$$d_{k,j} = F''(x_{k,j})^{-\frac{1}{2}} d'_{k,j} \tag{4.16}$$

and define the predicted reduction in the merit function (4.1) by

$$Pred_{k,j} = m_{k,j}(0) - m_{k,j}(d_{k,j}) = -q'_{k,j}(d'_{k,j}). \tag{4.17}$$

The feasible set is denoted by $F_p = \{x \in E | Ax = b, x \in K\}$. Now we are ready to present our algorithm.

ALGORITHM 4.1 (an interior-point trust-region algorithm).

Step 0 **Initialization.** *An initial point $x_{0,0} \in ri\{F_p\}$, an initial trust-region radius $\alpha_{0,0} \in (0,1)$, and an initial parameter $\eta_0 > 0$ are given. The constants η'_1, η'_2, γ_1, and γ_2 are also given and satisfy $0 < \eta'_1 \le \eta'_2 < 1$ and $0 < \gamma_1 \le \gamma_2 < 1$. Two tolerance numbers $\epsilon_1, \epsilon_2 \in (0,1)$ are given. Set $k = 0$ and $j = 0$.*

Step 1 **Test inner iteration termination.** *If $\eta_k\|p_{k,j}\| < \epsilon_1$ and $\eta_k\lambda_{k,j} > -\epsilon_2$, set $x_{k+1,0} = x_{k,j}$ and go to Step* 5.

Step 2 **Step calculation.** *Solve* (4.11)–(4.13) *to obtain $d'_{k,j}$, which satisfies* (4.14) *and* (4.15)*, and set $d_{k,j}$ by* (4.16).

Step 3 **Acceptance of the trial point.** *If $x_{k,j} + d_{k,j} \notin ri\{F_p\}$, set $\rho_{k,j} = -\infty, x_{k,j+1} = x_{k,j}$ and go to Step* 4*; otherwise compute the ratio*

$$\rho_{k,j} = \frac{f_{\eta_k}(x_{k,j}) - f_{\eta_k}(x_{k,j} + d_{k,j})}{Pred_{k,j}}. \tag{4.18}$$

Let

$$x_{k,j+1} = \begin{cases} x_{k,j} + d_{k,j} & \text{if } \rho_{k,j} \geq \eta_1', \\ x_{k,j} & \text{otherwise.} \end{cases} \tag{4.19}$$

Step 4 **Trust-region radius update.** *If* $\rho_{k,j} \geq \eta_2'$, *set* $\alpha_{k,j+1} \in [\alpha_{k,j}, \infty)$; *if* $\rho_{k,j} \geq \eta_1'$, *set* $\alpha_{k,j+1} \in [\gamma_2 \alpha_{k,j}, \alpha_{k,j}]$; *if* $\rho_{k,j} < \eta_1'$, *set* $\alpha_{k,j+1} \in [\gamma_1 \alpha_{k,j}, \gamma_2 \alpha_{k,j}]$; *increase* j *by* 1 *and go to Step* 1.

Step 5 **Update parameter** η. *Choose* $\eta_{k+1} > \eta_k$ *in such a way as to ensure that* $\eta_k \to +\infty$ *when* $k \to +\infty$. *Increase* k *by* 1 *and go to Step* 1.

We want to mention that from Lemma 2.1, our trail point will always stay inside the feasible set if we keep $\alpha_{k,j}$ less than 1. However, we believe that our current mechanism can make the algorithm more efficient without keeping $\alpha_{k,j}$ less than 1. Although it allows the feasibility to not be achieved, sufficient descent of the merit function can be achieved in a successful step. Just like the usual notation in the trust-region literature, if $\rho_{k,j} \geq \eta_2'$, we call this iteration very successful; if $\rho_{k,j} \geq \eta_1'$, we call this iteration successful; if $\rho_{k,j} < \eta_1'$, we call this iteration a failure. Since $p_{k,j}$ is the projection of $c_{k,j}$ onto the null space of $A_{k,j}$, there exists a vector $y \in R^m$ such that

$$p_{k,j} = c_{k,j} - (A_{k,j})^* y. \tag{4.20}$$

LEMMA 4.1. *Let* $p_{k,j}$ *be given by* (4.20) *and*

$$s_{k,j} = Q x_{k,j} + c - A^* y; \tag{4.21}$$

if $\eta_k \|p_{k,j}\| < 1$, *then* $s_{k,j} \in K^\circ$ *and*

$$\langle x_{k,j}, s_{k,j} \rangle \leq \frac{1}{\eta_k}(\sqrt{\vartheta} + \vartheta). \tag{4.22}$$

Proof. It follows from (4.20) that

$$\begin{aligned} p_{k,j} = c_{k,j} - (A_{k,j})^* y &= F''(x_{k,j})^{-\frac{1}{2}} Q x_{k,j} + F''(x_{k,j})^{-\frac{1}{2}} c \\ &\quad + \frac{1}{\eta_k} F''(x_{k,j})^{-\frac{1}{2}} F'(x_{k,j}) - (A F''(x_{k,j})^{-\frac{1}{2}})^* y \\ &= F''(x_{k,j})^{-\frac{1}{2}} s_{k,j} + \frac{1}{\eta_k} F''(x_{k,j})^{-\frac{1}{2}} F'(x_{k,j}). \end{aligned} \tag{4.23}$$

Therefore, the above relation and our assumption $\eta_k \|p_{k,j}\| < 1$ imply that

$$\begin{aligned} \|F''(x_{k,j})^{-\frac{1}{2}}(\eta_k s_{k,j} + F'(x_{k,j}))\| &= \|F''(x_{k,j})^{\frac{1}{2}}(F''(x_{k,j})^{-1} \eta_k s_{k,j} - x_{k,j})\| \\ &< 1. \end{aligned} \tag{4.24}$$

Here the last equality follows from Lemma 2.3. Then from Lemma 2.1, we know that $F''(x_{k,j})^{-1} \eta_k s_{k,j} \in K^\circ$. It follows from Lemma 2.4 that $\eta_k s_{k,j} \in K^\circ$, and consequently $s_{k,j} \in K^\circ$.

It follows from (4.23) that $s_{k,j} = \frac{1}{\eta_k}(F''(x_{k,j})^{\frac{1}{2}}(\eta_k p_{k,j}) - F'(x_{k,j}))$. Consequently, we have that

$$\begin{aligned}
\langle x_{k,j}, s_{k,j}\rangle &= \frac{1}{\eta_k}(\langle F''(x_{k,j})^{\frac{1}{2}} x_{k,j}, \eta_k p_{k,j}\rangle + \langle x_{k,j}, -F'(x_{k,j})\rangle) \\
&\le \frac{1}{\eta_k}(\|F''(x_{k,j})^{\frac{1}{2}} x_{k,j}\|\|\eta_k p_{k,j}\| + \langle x_{k,j}, -F'(x_{k,j})\rangle) \\
&\le \frac{1}{\eta_k}(\langle x_{k,j}, F''(x_{k,j}) x_{k,j}\rangle^{\frac{1}{2}} + \langle x_{k,j}, -F'(x_{k,j})\rangle) \\
&= \frac{1}{\eta_k}(\langle x_{k,j}, -F'(x_{k,j})\rangle^{\frac{1}{2}} + \langle x_{k,j}, -F'(x_{k,j})\rangle) \\
&= \frac{1}{\eta_k}(\sqrt{\vartheta} + \vartheta). \qquad (4.25)
\end{aligned}$$

The last two equalities follow from Lemma 2.3. □

This convergence estimate is remarkable, considering the problem is nonconvex. We can achieve this mainly due to the special properties of the *ϑ-normal barrier*. From this estimate, we can see that the barrier parameter ϑ determines the complexity of our problem, which coincides with its role in the interior-point algorithm for convex programming.

For the rest of this section, we will show that the stop rule for the inner iterations can be satisfied in finitely many iterations. First, the following two lemmas are indispensable for our analysis.

LEMMA 4.2. (a) *The map $x \mapsto F''(x)^{-\frac{1}{2}}$ is continuous on the feasible set $\mathcal{F}_p$.*

(b) *There is a constant $C > 0$, such that $\|F''(x)^{-\frac{1}{2}}\| \le C$ for any $x \in \mathcal{F}_p$.*

For the positive orthant case, $F''(x)^{-\frac{1}{2}} = X = \mathrm{diag}\{x_1, x_2, \ldots, x_n\}$, and for the cone of semidefinite matrices, $F''(X)^{-\frac{1}{2}}\xi = X^{\frac{1}{2}}\xi X^{\frac{1}{2}}$. Therefore, part (a) is obviously true for these two cases. For the general symmetric cone, it is still true from the Jordan algebra point of view. We will give an explanation in the appendix. Part (b) follows immediately form part (a) and our assumption that $\mathcal{F}_p$ is bounded.

LEMMA 4.3. *There exists a positive constant C' such that if*

$$\alpha_{k,j} \le \min\left\{\frac{(1-\eta_2')\theta\eta_k\|p_{k,j}\|}{1+(1-\eta_2')\theta\eta_k\|p_{k,j}\|}, \frac{\|p_{k,j\|}}{C'}\right\}, \qquad (4.26)$$

then the iteration $\{k, j\}$ is very successful and $\alpha_{k,j+1} \ge \alpha_{k,j}$.

Proof. Let $C' = 1 + C^2\|Q\|$, where C is defined as in Lemma 4.2. It follows from the definition of $\beta_{k,j}$ and the last lemma that

$$\beta_{k,j} = 1 + \|Q_{k,j}\| \le C'. \qquad (4.27)$$

Therefore, when $\alpha_{k,j} \le \frac{\|p_{k,j}\|}{C'}$, the inequality (4.14) becomes

$$q'(d'_{k,j}) \le -\theta\|p_{k,j}\|\alpha_{k,j}. \qquad (4.28)$$

Inequality $\alpha_{k,j} \le \frac{(1-\eta_2')\theta\eta_k\|p_{k,j}\|}{1+(1-\eta_2')\theta\eta_k\|p_{k,j}\|} < 1$ ensures that $x_{k,j} + d_{k,j} \in ri\{\mathcal{F}_p\}$. It follows

from inequalities (4.3) and (4.28) that

$$|\rho_{k,j}-1| = \left|\frac{f_{\eta_k}(x_{k,j}) - f_{\eta_k}(x_{k,j}+d_{k,j}) + m_{k,j}(d_{k,j})}{Pred_{k,j}}\right| \le \frac{\frac{(\alpha_{k,j})^3}{3(1-\alpha_{k,j})}}{\theta\alpha_{k,j}\eta_k\|p_{k,j}\|}$$

$$< \frac{\frac{\alpha_{k,j}}{1-\alpha_{k,j}}}{\theta\eta_k\|p_{k,j}\|} \le 1-\eta_2'. \tag{4.29}$$

Therefore, $-(\rho_{k,j}-1) \le |\rho_{k,j}-1| \le 1-\eta_2'$, and we can see that $\rho_{k,j} \ge \eta_2'$. Consequently, the iteration is very successful and $\alpha_{k,j+1} \ge \alpha_{k,j}$. □

LEMMA 4.4. *If $\eta_k\|p_{k,j}\| \ge \epsilon$ for some constant $\epsilon \in (0, \alpha_{k,0})$ and all j, then*

$$\alpha_{k,j} \ge \min\left\{\frac{\gamma_1(1-\eta_2')\theta\epsilon}{1+(1-\eta_2')\theta\epsilon}, \frac{\gamma_1\epsilon}{C'\eta_k}\right\} \tag{4.30}$$

holds for all j.

Proof. It is easy to see that (4.30) holds for $j=0$ as $\alpha_{k,0} > \epsilon$. Assume that the j is the first integer such that $\alpha_{k,j+1} < \min\{\frac{\gamma_1(1-\eta_2')\epsilon}{1+(1-\eta_2')\epsilon}, \frac{\gamma_1\epsilon}{C'\eta_k}\}$; from the update of the trust-region radius, we know $\gamma_1\alpha_{k,j} \le \alpha_{k,j+1}$, and hence

$$\alpha_{k,j} < \min\left\{\frac{(1-\eta_2')\theta\epsilon}{1+(1-\eta_2')\theta\epsilon}, \frac{\epsilon}{C'\eta_k}\right\}$$

$$< \min\left\{\frac{(1-\eta_2')\theta\eta_k\|p_{k,j}\|}{1+(1-\eta_2')\theta\eta_k\|p_{k,j}\|}, \frac{\|p_{k,j\|}}{C'}\right\}. \tag{4.31}$$

From the above inequality and the last lemma, we have that $\alpha_{k,j+1} \ge \alpha_{k,j}$, which contradicts the assumption that $\alpha_{k,j+1}$ is the first trust-region radius violating (4.30). The contradiction shows that the lemma is true. □

Now we are ready to prove that the first part of the stopping rule, i.e., $\eta_k\|p_{k,j}\| < \epsilon_1$, can be satisfied in finitely many iterations.

LEMMA 4.5. (a) *If there are only finitely many successful iterations in each inner algorithm, then $x_{k,j} = x^*$ and $\|p_{k,j}\| = \|p(x^*)\| = 0$ for all sufficiently large j.*

(b) $\liminf_{j\to\infty} \eta_k\|p_{k,j}\| = 0$.

(c) $\lim_{j\to\infty} \eta_k\|p_{k,j}\| = 0$.

Proof. (a) The mechanism of the algorithm ensures that $x^* = x_{k,j_0} = x_{k,j}$ for all $j > j_0$, where $\{k, j_0\}$ is the index of the last successful iterate. Since all iterations are unsuccessful for sufficiently large j, we know $\alpha_{k,j}$ will converge to zero. If $\|p(x^*)\| = \|p_{k,j_0}\| > 0$, Lemma 4.4 implies that $\alpha_{k,j}$ will be bounded from zero. This contradiction shows that $\|p_{k,j_0}\|$ has to be zero.

(b) For the purpose of deriving contradiction, we assume that for all j, $\eta_k\|p_{k,j}\| \ge \epsilon$ for some $\epsilon > 0$. From Lemma 4.4 we know that $\alpha_{k,j} \ge \min\{\frac{\gamma_1(1-\eta_2)\theta\epsilon}{1+(1-\eta_2)\theta\epsilon}, \frac{\gamma_1\epsilon}{C'\eta_k}\}$ for all j. We consider all successful iterations $\{k, j\}$; then

$$f_{\eta_k}(x_{k,j}) - f_{\eta_k}(x_{k,j}+d_{x_{k,j}}) \ge \eta_1' Pred_{k,j}$$

$$\ge \eta_1'\theta\|p_{k,j}\| \min\left\{\frac{\|p_{k,j}\|}{\beta_{k,j}}, \alpha_{k,j}\right\}; \tag{4.32}$$

here the last inequality follows by inequality (4.14). From the above analysis, we know $\eta_1'\theta\|p_{k,j}\| \min\{\frac{\|p_{k,j}\|}{\beta_{k,j}}, \alpha_{k,j}\} \ge \sigma > 0$; here σ is some positive constant number

that is independent of j. If we have infinitely many successful iterations, the difference between $f_{\eta_k}(x_{k,0})$ and $f_{\eta_k}(x_{k,j})$ will be unbounded when $j \to +\infty$. This contradicts the assumption that $f_{\eta_k}(x)$ is bounded from below on the feasible set. Hence, we conclude that $\liminf_{j\to\infty} \eta_k \|p_{k,j}\| = 0$.

(c) For the purpose of deriving a contradiction, assume there is a subsequence of successful iterations $\{x_{k,j_i}\}$ such that $\eta_k\|p_{k,j_i}\| \geq 2\epsilon$ for some $\epsilon > 0$ and for all $\{j_i\}$. Our part (b) ensures the existence for each $\{j_i\}$ of a first successful iteration $l_i = l(j_i) > j_i$ such that $\eta_k\|p_{k,l_i}\| < \epsilon$. We thus obtain another subsequence of successful iterations $\{l_i\}$ such that $\eta_k\|p_{k,j}\| \geq \epsilon$ for $j_i \leq j < l_i$ and $\eta_k\|p_{k,l_i}\| < \epsilon$. Define $\kappa = \{j \in S | j_i \leq j < l_i\}$; here S indicates the successful iterations. For $j \in \kappa$, from inequality (4.32) we have

$$\begin{aligned} f_{\eta_k}(x_{k,j}) - f_{\eta_k}(x_{k,j+1}) &\geq \eta_1'\theta\|p_{k,j}\| \min\left\{\frac{\|p_{k,j}\|}{\beta_{k,j}}, \alpha_{k,j}\right\} \\ &\geq \eta_1'\theta\frac{\epsilon}{\eta_k}\min\left\{\frac{\epsilon}{\eta_k\beta_{k,j}}, \alpha_{k,j}\right\}. \end{aligned} \tag{4.33}$$

Since the sequence $f_{\eta_k}(x_{k,j})$ is monotonically decreasing and bounded from below, it is convergent. Therefore, the left-hand side of (4.33) must tend to zero when j tends to infinity. This gives that $\lim_{j\to\infty, j\in\kappa} \alpha_{k,j} = 0$. As a consequence, the second term dominates the minimum in (4.33) and we obtain that for $j \in \kappa$ sufficiently large,

$$\alpha_{k,j} \leq \frac{2\eta_k(f_{\eta_k}(x_{k,j}) - f_{\eta_k}(x_{k,j+1}))}{\eta_1'\theta\epsilon}. \tag{4.34}$$

We then deduce from this bound that, for i sufficiently large,

$$\begin{aligned} \|x_{k,j_i} - x_{k,l_i}\| &\leq \sum_{j=j_i, j\in\kappa}^{l_i-1} \|d_{x_{k,j}}\| = \sum_{j=j_i, j\in\kappa}^{l_i-1} \|F''(x_{k,j})^{-\frac{1}{2}} d'_{x_{k,j}}\| \\ &\leq \sum_{j=j_i, j\in\kappa}^{l_i-1} \|F''(x_{k,j})^{-\frac{1}{2}}\| \|d'_{x_{k,j}}\| \leq \sum_{j=j_i, j\in\kappa}^{l_i-1} C\alpha_{k,j} \\ &\leq C \sum_{j=j_i, j\in\kappa}^{l_i-1} \frac{2\eta_k(f_{\eta_k}(x_{k,j}) - f_{\eta_k}(x_{k,j+1}))}{\eta_1'\theta\epsilon} \\ &= \frac{2C\eta_k(f_{\eta_k}(x_{k,j_i}) - f_{\eta_k}(x_{k,l_i}))}{\eta_1'\theta\epsilon}. \end{aligned} \tag{4.35}$$

Here the third inequality follows from part (b) of Lemma 4.2, and the fourth inequality follows from inequality (4.34). Because $f_{\eta_k}(x_{k,j})$ is monotonically decreasing for j and bounded from below, it is convergent. Consequently, $f_{\eta_k}(x_{k,j_i}) - f_{\eta_k}(x_{k,l_i})$ tends to zero when $i \to +\infty$. We therefore obtain that $\|x_{k,j_i} - x_{k,l_i}\|$ tends to zero when $i \to +\infty$. Without loss of generality, we can assume x^* to be the common limit point of sequences $\{x_{k,j_i}\}_{i=1}^{\infty}$ and $\{x_{k,l_i}\}_{i=1}^{\infty}$. Since the sequences of our algorithm make the value of $f_{\eta_k}(x)$ decrease, the limit point x^* must be in the interior of the feasible set F_p. We know

$$\|p_{k,j_i} - p_{k,l_i}\| \leq \|p_{k,j_i} - p(x^*)\| + \|p(x^*) - p_{k,l_i}\|. \tag{4.36}$$

From part (a) of Lemma 4.2, we know that $c_{k,j}$ is continuous for x. Since $p_{k,j}$ is the projection of $c_{k,j}$ over the null space of $A_{k,j}$, $p_{k,j}$ is also continuous for x. Therefore, the right-hand side of inequality (4.36) will converge to zero when i tends to infinity. But, on the other hand, we know $\eta_k\|p_{k,j_i} - p_{k,l_i}\| \geq \eta_k\|p_{k,j_i}\| - \eta_k\|p_{k,l_i}\| \geq \epsilon$. Therefore, we get a contradiction, which means our initial assumption that $\eta_k\|p_{k,j}\|$ does not converge to zero cannot be true. This completes our proof. ∎

Now we prove that the second part of the stop rule, $\eta_k\lambda_{k,j} > -\epsilon_2$, can also be satisfied in finitely many iterations.

LEMMA 4.6. *For every fixed k, $\limsup_{j\to\infty} \lambda_{k,j} \geq 0$.*

Proof. For the purpose of deriving a contradiction, we assume that $\lambda_{k,j} \leq \lambda_*$ for some $\lambda_* < 0$ and all j. From inequality (4.15) we know that

$$Pred_{k,j} = -q'(d'_{x_{k,j}}) \geq -\theta\lambda_{k,j}\min\{\lambda_{k,j}^2, \alpha_{k,j}^2\} \geq -\theta\lambda_*\min\{\lambda_*^2, \alpha_{k,j}^2\}. \tag{4.37}$$

Therefore, we get

$$|\rho_{k,j} - 1| = \left|\frac{f_{\eta_k}(x_{k,j}) - f_{\eta_k}(x_{k,j} + d_{k,j}) + m_{k,j}(d_{k,j})}{Pred_{k,j}}\right| \leq \frac{\frac{(\alpha_{k,j})^3}{3(1-\alpha_{k,j})}}{-\eta_k\theta\lambda_*\min\{\lambda_*^2, \alpha_{k,j}^2\}}. \tag{4.38}$$

From this inequality, there exists a constant $\delta_1 > 0$ such that if $\alpha_{k,j} < \delta_1$, then $|\rho_{k,j} - 1| \leq 1 - \eta_2'$, that is, $\rho_{k,j} \geq \eta_2'$, which means this iteration is very successful and $\alpha_{k,j+1} \geq \alpha_{k,j}$. Now we assume $\{k, j_0\}$ is the first iteration such that $\alpha_{k,j_0} \leq \delta_1$; then from our above analysis, we know that $\alpha_{k,j} \geq \min\{\gamma_1\delta_1, \alpha_{k,j_0}\} := \delta_2$ for all $j \geq j_0$. Consequently,

$$\begin{aligned} f_{\eta_k}(x_{k,j}) - f_{\eta_k}(x_{k,j} + d_{x_{k,j}}) &\geq \eta_1' Pred_{k,j} \geq -\eta_1'\theta\lambda_*\min\{\lambda_*^2, \alpha_{k,j}^2\} \\ &\geq -\eta_1'\theta\lambda_*\min\{\lambda_*^2, \delta_2^2\} > 0 \end{aligned} \tag{4.39}$$

whenever $\{k, j\}$ is successful. If there are infinitely many successful iterations after $\{k, j_0\}$, (4.39) contradicts the fact that $f_{\eta_k}(x)$ is bounded from below. If there are finitely many successful iterations, the mechanism of our algorithm ensures that $\alpha_{k,j}$ converges to zero. But it again contradicts $\alpha_{k,j} \geq \min\{\gamma_1\delta_1, \alpha_{k,j_0}\} := \delta_2$ for all $j \geq j_0$. Hence our original assumption that there exists $\lambda_* < 0$ such that for all j, $\lambda_{k,j} \leq \lambda_*$ cannot be true. This completes the proof. ∎

From Lemma 4.6 and part (c) of Lemma 4.5, it is obvious that the stopping rule of our inner algorithm can be satisfied in finite many iterations.

THEOREM 4.1. *For every fixed η_k, Step 1 through Step 4 can be terminated in finitely many iterations.*

Finally, we can derive that any limit point of the sequences our algorithm generates satisfies both the first-order and the second-order optimality conditions.

THEOREM 4.2. *Assume x^* is any limit point of the sequences $\{x_{k,0}\}_{k=0}^{\infty}$ our algorithm generates; then x^* satisfies both the first-order and the second-order optimality conditions for our problem* (3.1)–(3.3).

Proof. We assume that $s_{k+1,0}$ is defined by (4.21). From Lemma 4.1, we know

$$\langle x_{k+1,0}, s_{k+1,0}\rangle \leq \frac{1}{\eta_k}(\vartheta + \sqrt{\vartheta}). \tag{4.40}$$

The above inequality implies that $\eta_k \to \infty$ as $k \to \infty$. Hence, any limit point of the sequences $\{x_{k,0}\}_{k=0}^{\infty}$ must satisfy the first-order optimality condition. Moreover, we know that

$$\lambda_{k+1,0} \geq \frac{-\epsilon_2}{\eta_k}, \tag{4.41}$$

which implies that

$$\liminf_{k\to\infty} \lambda_{k,0} \geq 0. \tag{4.42}$$

The above inequality, the definition of $\lambda_{k,j}$, and our continuity assumption show that $F''(x^*)^{-\frac{1}{2}} Q F''(x^*)^{-\frac{1}{2}}$ is positive semidefinite on the vector space $\{x | AF''(x^*)^{-\frac{1}{2}} x = 0, x \in E\}$. From Corollary 3.1, we know x^* satisfies the second-order optimality condition. □

5. Solve the large-scale trust-region subproblem. In this section, we show how to use our algorithm to solve the trust-region subproblem exactly and approximately. Numerical results are presented.

Consider the following standard trust-region subproblem:

$$\min \quad q(x) = \frac{1}{2}\langle x, Qx\rangle + \langle c, x\rangle \tag{5.1}$$

$$\text{subject to} \quad \|x\| \leq \Delta, \tag{5.2}$$

where $\|.\|$ is the ℓ_2-norm. By introducing a new variable x_{n+1}, we can transform this problem into the following nonlinear second-order cone programming:

$$\min \quad q(x) = \frac{1}{2}\langle x, Qx\rangle + \langle c, x\rangle \tag{5.3}$$

$$\text{subject to} \quad x_{n+1} = \Delta, \tag{5.4}$$

$$\sum_{i=1}^{n} x_i^2 \leq x_{n+1}^2. \tag{5.5}$$

Obviously, this is a special symmetric cone programming with $A = (0 \cdots 0, 1)$ and $K = \{x \in R^{n+1} : \sum_{i=1}^{n} x_i^2 \leq x_{n+1}^2 \text{ and } x_{n+1} \geq 0\}$. If we want to use Algorithm 4.1 to solve (5.3)–(5.5), we need to choose a method of solving (4.11)–(4.13). Since we are interested in solving large-scale problems, this motivates us to choose the methods that rely only on matrix-vector product. The first method in this class is the Steihaug–Toint truncated conjugate gradient method, which is due to Toint [26] and Steihaug [23]. And the adaptation to handle additional affine constraints can be found in Gould, Hribar, and Nocedal [15]. Here we give the version of the algorithm for solving (4.11)–(4.13).

ALGORITHM 5.1 (the Steihaug–Toint method with affine constraints).

Step 0 **Initialization.** *For fixed* $\{k, j\}$ *in* (4.11)–(4.13), *let* $d'_0 = 0$, $g_0 = c_{k,j}$, $v_0 = P_{A_{k,j}} c_{k,j}$, *and* $p_0 = -v_0$. *For* $h = 0, 1, \ldots$ *until convergence, perform the iteration.*

Step 1 **Check the negative curvature.** *Set* $\kappa_h = \langle p_h, Q_{k,j} p_h\rangle$. *If* $\kappa_h \leq 0$, *compute* σ_h *as the positive root of* $\|d'_h + \sigma p_h\| = \alpha_{k,j}$, *set* $d'_{h+1} = d'_h + \sigma_h p_h$, *and stop. End if*

Step 2 **Check the boundary constraints.** *Set* $\beta_h = \langle g_h, v_h\rangle/\kappa_h$. *If* $\|d'_h + \beta_h p_h\| \geq \alpha_{k,j}$, *compute* σ_h *as the positive root of* $\|d'_h + \sigma p_h\| = \alpha_{k,j}$, *set* $d'_{h+1} = d'_h + \sigma_h p_h$, *and stop. End if*

Step 3 **Perform the conjugate gradient step.** *Set* $d'_{h+1} = d'_h + \beta_h p_h$, $g_{h+1} = g_h + \beta_h Q_{k,j} p_h$, $v_{h+1} = P_{A_{k,j}} g_{h+1}$, *and* $p_{h+1} = -v_{h+1} + \frac{\langle g_{h+1}, v_{h+1}\rangle}{\langle g_h, v_h\rangle} p_h$.

Here $P_{A_{k,j}}$ is the projection onto the null space of $A_{k,j}$.

There are several advantages of this algorithm:

(1) It requires only matrix-vector product.
(2) It usually terminates very fast.
(3) It is applicable to problems with affine constraints.
(4) If the objective function is convex, the computed approximation solution gives at least half of the optimal reduction (Yuan [32]).

This Steihaug–Toint method is basically unconcerned with the trust region until it blunders into its boundary and stops. This is rather unfortunate, particularly, as considerable experience has shown that this frequently happens during the first few iterations when a negative curvature is present, causing the following disadvantages to the algorithm:

(A) Even if the problem is convex, optimal solution cannot be expected, except when the solution lies interior to the trust region.
(B) If it blunders into the boundary or a negative curvature is present too early, the approximate solution is not very good.
(C) It cannot handle the hard case.
(D) Optimal solution for the nonconvex problem is normally impossible for this algorithm.

Can we remove these disadvantages of Algorithm 5.1 while retaining its advantages? The answer is yes. After transforming (5.1)–(5.2) into (5.3)–(5.5), we use Algorithm 4.1 to solve it. In each iteration we use the Steihaug–Toint conjugate gradient method to solve (4.11)–(4.13). Since we basically repeat using Algorithm 5.1 in each iteration, we can keep all the advantages of Algorithm 5.1 as long as the number of iterations is not too big. It turns out that the number of iterations is very reasonable from the numerical results presented in this section. What can we achieve by doing this? We can get at least a first-order critical point of (5.3)–(5.5). This gives an optimal solution of (5.1)–(5.2) if Q is positive semidefinite. Thus we have removed (A). Algorithm 5.1 sometimes cannot give us a good approximate solution because it hits the boundary too early. By adding a ϑ-norm barrier to the quadratic model, we can prevent the iterates to reach the boundary too soon. This idea can give us a much better approximate solution, which is verified by the numerical results in this section. Therefore, we have removed (B). We know that Algorithm 5.1 cannot handle the hard case. If $c = 0$ and Q is indefinite, the method will terminate at $d' = 0$ with no decrease in the model. This cannot happen in our new framework. For Algorithm 4.1, a first-order critical point is always ensured even if the problem is in the hard case. Therefore, we have removed (C). Moreover, Algorithm 4.1 can be improved to find the optimal solution of (5.1)–(5.2) for all the cases, including the nonconvex case and the hard case. We first need the following lemma, which is well known in the trust-region literature.

LEMMA 5.1. *Any global minimizer* x^* *of* (5.1)–(5.2) *satisfies the equation*

$$(Q + \mu^* I)x^* = -c; \tag{5.6}$$

here $Q + \mu^* I$ *is positive semidefinite,* $\mu^* \geq 0$, *and* $\mu^*(\|x^*\| - \Delta) = 0$.

For a proof, see, e.g., Section 7.2 of Conn, Gould, and Toint [6].

The following algorithm removes (D).

ALGORITHM 5.2 (an algorithm for optimal solution).

Step 0 **Make Q positive semidefinite** *Find s, the smallest eigenvalue of Q and v, its corresponding eigenvector. If $s < 0$, set $Q = Q - sI$, end if.*

Step 1 **Solve new model.** *Use Algorithm* 4.1 *to solve* (5.3)–(5.5) *with Q positive semidefinite to get solution x_0.*

Step 2 **Go back to original model.** *If $s \geq 0$, set $x = x_0$, end if. If $s < 0$ and $\|x_0\| = \Delta$, set $x = x_0$, end if. If $s < 0$ and $\|x_0\| < \Delta$, set $x = x_0 + \sigma v$, σ is chosen so that $\|x_0 + \sigma v\| = \Delta$, end if.*

We claim that x is an optimal solution of (5.1)–(5.2). If $s \geq 0$, it is obvious. If $s < 0$ and $\|x_0\| = \Delta$, it follows from the fact that x_0 is an optimal solution of the new model with Q positive semidefinite and Lemma 5.1. If $s < 0$ and $\|x_0\| < \Delta$, $\mu^* = 0$ in (5.6) of Lemma 5.1 and consequently $Qx_0 = -c$ for the new convex Q. And since $Qv = 0$ for the new Q, $Qx = Q(x_0 + \sigma v) = -c$. From Lemma 5.1, we know that x is an optimal solution of (5.1)–(5.2). Therefore, we have removed all the disadvantages of the Steihaug–Toint method, while our algorithms mainly rely on the conjugate gradient method. For finding the optimal solution of the problem when it is nonconvex, we need to compute the least eigenvalue. However, those eigenvalue-based algorithms like those of Sorensen [22], Rojas, Santos, and Sorensen [21], and Rendl and Wolkowicz [19] require computing sequences of least eigenvalues, while we compute the least eigenvalue only once. As pointed out to us by a referee, Griffin and Gill [16] independently applied a truncated conjugate gradient algorithm to a shifted quadratic function to solve the trust-region subproblem.

We need to mention two implementation techniques when we use Algorithm 5.1 to solve (4.11)–(4.13) in each iteration.

We can see that the main computation in this algorithm is the product of the matrix $Q_{k,j}$ with a vector. In practice, we do not form $Q_{k,j}$ explicitly because it is expensive and destroys the sparse structure of Q. Since $Q_{k,j} = F''(x_{k,j})^{-\frac{1}{2}} Q F''(x_{k,j})^{-\frac{1}{2}} + \frac{1}{\eta_k} I$, we need to compute the product of the matrix $F''(x_{k,j})^{-\frac{1}{2}}$ with a vector, which can be done efficiently. For $F(x) = -\ln(x_{n+1}^2 - \sum_{i=1}^n x_i^2)$, $F''(x)^{-1}$ and $F''(x)^{-\frac{1}{2}}$ have the following explicit forms:

$$F''(x)^{-1} = \frac{1}{2}\begin{bmatrix} (x_{n+1}^2 - y^T y)I + 2yy^T & 2x_{n+1}y \\ 2x_{n+1}y^T & x_{n+1}^2 + y^T y \end{bmatrix}, \tag{5.7}$$

$$F''(x)^{-\frac{1}{2}} = \frac{\sqrt{2}}{2}\begin{bmatrix} \sqrt{x_{n+1}^2 - y^T y}I + \frac{yy^T}{\sqrt{x_{n+1}^2 - y^T y} + x_{n+1}} & y \\ y^T & x_{n+1} \end{bmatrix}, \tag{5.8}$$

where $y = (x_1 \dots x_n)^T$. For more details about the second-order cone and its barrier, see, e.g., Alizadeh and Goldfarb [2] or Faybusovich and Tsuchiya [10].

The first technique is that we do not have to formulate $F''(x)^{-\frac{1}{2}}$ explicitly for computing the product of the matrix $F''(x_{k,j})^{-\frac{1}{2}}$ with a vector. Since $yy^T u = \langle y, u\rangle y$ for any vector $u \in R^n$, the computation needs only $O(n)$ arithmetic operations.

The second technique is for the projection $P_{A_{k,j}}$. For any vector $u \in R^{n+1}$,

$$\begin{aligned} P_{A_{k,j}} u &= u - A_{k,j}^T (A_{k,j} A_{k,j}^T)^{-1} A_{k,j} u \\ &= u - F''(x_{k,j})^{-\frac{1}{2}} A^T (A F''(x_{k,j})^{-1} A^T)^{-1} A F''(x_{k,j})^{-\frac{1}{2}} u. \end{aligned} \tag{5.9}$$

Because $A = (0 \dots 0, 1)$, $AF''(x_{k,j})^{-1}A^T$ is just the $(n+1, n+1)$ entry of $F''(x_{k,j})^{-1}$. The only thing left is to compute the product of $F''(x_{k,j})^{-\frac{1}{2}}$ with a vector, which has been taken care of by $O(n)$ arithmetic operations.

By the above two techniques, each iteration of our algorithm takes hardly more extra work than the Steihaug–Toint truncated conjugate gradient method.

When applying Algorithm 4.1 to (5.3)–(5.5), we change the stopping rule for the inner iterations to make the algorithm more efficient. We remind the reader that there are two conditions of our stopping rule: (a) $\eta_k \|p_{k,j}\| < \epsilon_1$ and (b) $\eta_k \lambda_{k,j} > -\epsilon_2$ for some $\epsilon_1, \epsilon_2 \in (0, 1)$. We ignore condition (b), since a first-order critical point is good enough for our purpose. For condition (a), we need to change it a little because it is independent of whether or not optimality is nearly achieved. In practice, we directly follow the definition of the first-order optimality condition. If $x_{k,j}$ is a first-order critical point, $s_{k,j} = Qx_{k,j} + c - A^*y$ should be inside the second-order cone for some y. In our case $A = (0 \dots 0, 1)$ and all of the entries in the last row of Q are zeros. Therefore, we set $y = -\|Qx_{k,j} + c\|$ such that $s_{k,j}$ is inside the second-order cone. Then we stop the inner iteration as soon as we find $x_{k,j}$ such that $\langle x_{k,j}, s_{k,j} \rangle \le \frac{\epsilon}{\eta_k}$ for some constant ϵ. Lemma 4.1 suggests that $\epsilon = \sqrt{\vartheta} + \vartheta$ is a good choice. For the second-order cone, $\vartheta = 2$. From our practical experience, it works very well for convex problems. In the nonconvex case, it seems that this stopping rule works for some problems but not all of them, which remains to be investigated. The algorithm is halted as soon as $x_{k,j}$ is found such that $\langle x_{k,j}, s_{k,j} \rangle \le 10^{-4}$. In Algorithm 4.1, $\eta_1' = 0.05$ and $\eta_2' = 0.9$ are used and the trust region is updated according to the usual rule. If $\rho_{k,j} \ge \eta_2'$, set $\alpha_{k,j+1} = \max(\alpha_{k,j}, 2\|d_{k,j}'\|)$; if $\rho_{k,j} \in [\eta_1', \eta_2')$, set $\alpha_{k,j+1} = \alpha_{k,j}$; if $\rho_{k,j} < \eta_1'$, set $\alpha_{k,j+1} = \frac{1}{2}\alpha_{k,j}$. The initial value of parameter η is set to be $\frac{1}{\Delta}$ and is updated by $\eta_{k+1} = 10\eta_k$. In each iteration, the Steihaug–Toint conjugate gradient method (Algorithm 5.1) is stopped as soon as $\|v_h\| \le 10^{-\frac{3}{2}}\|v_0\|$ if it does not hit the boundary and negative curvature is not present before that. We decide not to put an upper bound on the number of Steihaug–Toint iterations, which is denoted by h in Algorithm 5.1. In this way, for the convex problems, we will be able to know the number of iterations our algorithm needs to get an optimal solution if we solve each trust-region subproblem approximately, which is measured by $\|v_h\| \le 10^{-\frac{3}{2}}\|v_0\|$.

The algorithms are tested in MATLAB 7.0 on a Linux system. We run our experiments on a Gateway computer with a Pentium IV 3.2G processor and 1G RAM. We compare our results with a software called "Newtrust4b" based on Rendel and Wolkowicz [19]. We choose "Newtrust4b" because it is one of the best software packages for finding the optimal solution of the trust-region subproblem and because it is also implemented in MATLAB code. For the test problems, Q and c are randomly generated with entries uniformly distributed on (0,1). We set the trust-region radius $\Delta = 1$. For different radiuses, the computation time and the number of iterations may vary, but they vary reasonably. In Tables 1–5, n is the dimension of the problems, d is the density of Q; i.e., Q has $d * n^2$ nonzero entries. The data in those columns under the algorithms' names is the computational time by seconds. Sometimes, the MATLAB timing is dependent on the CPU load. Our timings have been averaged to eliminate this dependency. And *its* is the number of iterations of our algorithm. ST *its* is the total number of Steihaug–Toint iterations used during the whole computation. From Algorithm 5.1, we can see that the dominating cost of each Steihaug–Toint iteration is the product of $Q_{k,j}$ and p_h. Therefore, ST *its* gives us the total number of the matrix-vector products used by our algorithm. We first test some convex problems. To make the problem convex, we let $Q = Q - sI$ if s, the least eigenvalue of

TABLE 1
Convex singular problems.

n	d	Newtrust4b	Algorithm 4.1	its	ST its
2500	1	57	3	10	66
5000	0.5	1328	67	28	156
10000	0.05	284	16	12	42
20000	0.01	240	10	9	19
100000	0.0001	181	5	8	15
200000	101 band	248	17	7	11

TABLE 2
Q is sparse with $d = 0.03$.

n	Newtrust4b	Algorithm 5.2	Eigenvalue time	its	ST its
4000	8	6	4	10	31
8000	74	68	58	22	53
12000	156	143	121	21	53
16000	303	278	239	21	54
20000	356	312	253	20	48

Q, is negative. In this way, the problem is convex and nearly singular. To show that our algorithm can handle the singular problems, we set $Q = \left[\begin{smallmatrix} Q & 0 \\ 0 & 0 \end{smallmatrix}\right]$ to make it even more singular. Now the dimension of Q is $n + 1$, and so is c. Since the problems are convex, Algorithm 4.1 gives us optimal solutions.

We can see that Algorithm 4.1 outperforms Newtrust4b for the convex singular problems. The success of Algorithm 4.1 in this case is a good basis for Algorithm 5.2. If the problem is nonconvex or if we do not know whether or not our problem is convex, we have to use Algorithm 5.2 to get an optimal solution. The following two groups of results show us the performance of Algorithm 5.2.

We can see that Algorithm 5.2 is competitive with Newtrust4b in both sparse and dense cases. The eigenvalue time is the computational time cost by computing the least eigenvalue of Q in Algorithm 5.2, which becomes dominant when the problem becomes large. Fortunately, we have reduced this part to the minimum level in Algorithm 5.2 (we need only compute the least eigenvalue once for all). Moreover, the number of iterations and ST iterations of our algorithm is independent of the dimension of the problems.

So far, we have been focusing on finding the optimal solution of the trust-region subproblem. But if the problem is nonconvex, Algorithm 4.1 can deliver us only an approximate solution. How good is Algorithm 4.1 for nonconvex problems? From our practical experience, we have to say that the performance of Algorithm 4.1 on finding an approximate solution for nonconvex problems is not as stable as its performance on finding exact solution for convex problems. This is reflected by the fact that the convergence is sensitive to the inner iteration stopping rule. The inner iteration stopping rule, $\langle x_{k,j}, s_{k,j}\rangle \le \frac{\sqrt{\vartheta}+\vartheta}{\eta_k}$, works for some of our testing problems but not for all of them. For those test problems where Algorithm 4.1 has good performance, the number of iterations is around 40. For those test problems where Algorithm 4.1 has bad performance, the number of iterations can reach over 100. The change of stopping rule of inner iteration (like from $\langle x_{k,j}, s_{k,j}\rangle \le \frac{\sqrt{\vartheta}+\vartheta}{\eta_k}$ to $\langle x_{k,j}, s_{k,j}\rangle \le \frac{1}{\eta_k}$) can significantly affect the performance of the algorithm on the same test problem. This phenomenon remains to be investigated. On the positive side, even for the problems where Algorithm 4.1 has bad performance, the convergence slows down only when η_k

TABLE 3
Q is dense with d = 1.

n	Newtrust4b	Algorithm 5.2	Eigenvalue time	its	ST its
1000	9	3	2	15	144
2000	14	7	4	22	141
3000	36	18	11	20	154
4000	147	71	56	22	144
5000	323	145	127	19	164

TABLE 4
Nonconvex singular problems.

n	d	Algorithm 4.1	Accuracy	its	ST its
3000	1	1	90%	5	10
6000	1	5	90%	5	10
10000	0.03	3	95%	3	6
20000	0.03	11	95%	3	6
100000	0.0001	2	99%	3	4
200000	101 band	8	99%	3	4

becomes large and our solution is close to the optimal solution. Here we give a group of examples to show that Algorithm 4.1 can deliver us a good approximate solution at a relatively low cost. For each of these problems, we first use Algorithm 5.2 to get an optimal solution and consequently the best possible reduction. Then we use Algorithm 4.1 to solve it and stop the algorithm when 90% of the optimal reduction is achieved. Q is randomly generated with entries uniformly distributed on (0,1). We have check that Q is indefinite. To show the performance of our algorithm on the singular problems, we set $Q = \left[\begin{smallmatrix} Q & 0 \\ 0 & 0 \end{smallmatrix}\right]$. Therefore, the actual dimension of Q in Table 4 is $n + 1$.

We can see that after a few Steihaug–Toint iterations, Algorithm 4.1 can deliver us a good approximate solution. Therefore, we have achieved our goal of improving the solution quality of the Steihaug–Toint method while keeping its computational advantages.

In summary, techniques developed in this section give us two algorithms for solving the trust-region subproblem. Algorithm 5.2 gives us an optimal solution for both convex and nonconvex problems. Algorithm 4.1 gives us a good approximate solution for nonconvex problems and an optimal solution for convex problems.

6. Further numerical results and implementation issues. In this section, we discuss some implementation issues for solving general symmetric cone programming. We also present some numerical results of solving a class of quadratic programming.

To solve the general symmetric cone programming, we have to handle three basic implementation issues. The first issue is to find a starting point in our feasible set. This feasible set has been well studied in the interior-point algorithm literature. We can use the same technique to find a feasible starting point for our problem. The second issue is to handle affine constraints. This requires us either to solve the normal equations or to project iterates onto the null space of $A_{k,j}$. Gould, Hribar, and Nocedal [15] is a good reference for handling this issue. The third issue is about preconditioning. We recall that $Q_{k,j} = F''(x_{k,j})^{-\frac{1}{2}} Q F''(x_{k,j})^{-\frac{1}{2}} + \frac{1}{\eta_k} I$. When we are getting close to the optimal solution, η_k is getting large, and consequently the right part $\frac{1}{\eta_k} I$ is about to disappear. At the same time, the iterate $x_{k,j}$ is getting close

TABLE 5
Q is positive definite with density d = 1.

n	its	ST its	Algorithm 4.1
1000	67	529	4
2000	75	799	17
3000	81	777	40
4000	71	630	75
5000	70	671	109

to the boundary. Therefore, $F''(x_{k,j})^{-\frac{1}{2}}$ becomes nearly singular, which can make the condition number of $Q_{k,j}$ large. As we know, the convergence behavior of the conjugate gradient method is strongly dependent on the conditioning of $Q_{k,j}$. Therefore, the appropriate preconditioning technique is necessary to make the algorithm efficient.

Handling these implementation issues is beyond the scope of this paper. However, to see how our algorithm performs on solving problems other than the trust-region subproblem, we present some numerical results for a class of quadratic programming, which is minimization of a strictly convex quadratic objective function over the positive orthant. This problem is bounded from below. We use the vector e with every entry 1 as our starting point. Q is randomly generated with entries uniformly distributed on (0,1). We make the problem convex by letting $Q = Q + (-s + 1)I$ if s, the least eigenvalue of Q, is negative. c is randomly generated with entries uniformly distributed on $(-1, 0)$. In this way, the problem will have nontrivial solution. In each step, the conjugate gradient method is stopped if the iterate hits the boundary or $\|g_h\| \leq 10^{-3/2}\|g_0\|$. The inner iteration is stopped when we find $x_{k,j}$ such that $s_{k,j} = Qx_{k,j} + c$ belongs to positive orthant and $\langle x_{k,j}, s_{k,j}\rangle \leq \frac{n+\sqrt{n}}{\eta_k}$. The algorithm is halted as soon as $x_{k,j}$ is found such that $\langle x_{k,j}, s_{k,j}\rangle \leq 10^{-4}$. All other implementation techniques are similar to those we discussed in section 5. The following is a group of results.

The number of iterations as well as the total number of Steihaug–Toint iterations are independent of the dimension of problems, which makes our algorithm have practical potential for solving large-scale problems. One practical observation we want to mention is that the computational time of reducing $\langle x_{k,j}, s_{k,j}\rangle$ from 10^{-3} to 10^{-4} is even more than the computational time of reducing it from the starting value to 10^{-3}. This is caused by the fact that the convergence of the conjugate gradient algorithm considerably slows down when the iterate is close to optimal solution and consequently the boundary, which agrees with our theoretical analysis above. Therefore, appropriate preconditioning is indispensable to make the algorithm more efficient. Since it has been shown in section 5 that Algorithm 4.1 works well for solving the singular problems, an alternative way is to use it to solve the trust-region subproblem when the iterate is close to the boundary of the positive orthant. Which way is better remains to be investigated.

7. Concluding remarks. In this paper, we have shown that combining the techniques developed in trust-region literature (especially Conn, Gould, and Toint [6]) with those techniques in interior-point method literature can be very powerful both in theoretical analysis and practical implementation. For further theoretical research, Lu and Yuan [17] have recently proved that the complexity of an interior-point trust-region algorithm for convex programming is polynomial time. On the practical side, the numerical results presented in this paper show that our algorithm

has practical potential. But a lot more work needs to be done to turn this method into a practical software package for solving general symmetric cone programming.

Appendix. In this appendix, we describe the face $V_{A'}$ for the semidefinite case. We use the Jordan algebra technique to prove Theorem 3.3 and give an explanation why Lemma 4.2 holds for general symmetric cone.

If $K = S_+^{n\times n}$, we know $F(X) = -\ln\det(X)$ is a *ϑ-normal barrier* for $S_+^{n\times n}$. Let $A' \in \partial K$, and $\operatorname{rank}(A') = r < n$; then $F''(A')^{-\frac{1}{2}}X = A'^{\frac{1}{2}}XA'^{\frac{1}{2}}$. We set $V = \{A'^{\frac{1}{2}}XA'^{\frac{1}{2}} | AA'^{\frac{1}{2}}XA'^{\frac{1}{2}} = 0, X \in S^{n\times n}\}$. In section 3, we have claimed that $V_{A'} = V$. Now we give a proof.

Proof. First, we need to characterize $V_{A'}$. Since A' is a semidefinite matrix, then we can find an orthogonal matrix U, such that $U^{-1}A'U = D$, where $D = \operatorname{diag}\{\lambda_1, \ldots, \lambda_r, 0, \ldots, 0\}$ with $\lambda_i > 0$, $i = 1, \ldots, r$.

Let $C = \operatorname{diag}\{0, \ldots, 0, 1, \ldots, 1\}$ be the matrix whose first r diagonal entries are 0 and the last $n - r$ diagonal entries are 1, and let $Q' = UCU^{-1}$. Then Q' is a nonzero positive semidefinite matrix and $\langle Q', A'\rangle = \langle UCU^{-1}, UDU^{-1}\rangle = \langle C, D\rangle = 0$. Furthermore, for any positive semidefinite $n \times n$ matrix X, $\langle Q', X\rangle \geq 0$. Therefore, the hyperplane $H = \{X \in S^{n\times n} | \langle Q', X\rangle = 0\}$ isolates $S_+^{n\times n}$ and contains A'. We claim that

$$\digamma_{A'} = \{X \in S_+^{n\times n} | AX = b, \langle Q', X\rangle = 0\}.$$

To see that A' is a relative interior point of $\digamma_{A'}$, we need only show that A' is an interior point of $\digamma = \{X \in S_+^{n\times n} | \langle Q', X\rangle = 0\}$. The map $X \mapsto Y = U^{-1}XU$ is a nondegenerate linear transform which maps $S_+^{n\times n}$ onto itself, maps Q' onto C, and maps A' onto D. Then $\digamma$ is mapped onto $\digamma' = \{Y \in S_+^{n\times n} | \langle C, Y\rangle = 0\}$. Clearly, Y must have the last $n - r$ rows and last $n - r$ columns be 0. The upper left $r \times r$ submatrix Y' of Y can be arbitrary positive semidefinite matrix, $Y = \left(\begin{smallmatrix} Y' & 0 \\ 0 & 0 \end{smallmatrix}\right)$. It is easy to see that $\digamma'$ contains D in its interior. Since $Y \mapsto X = UYU^{-1}$ is a nondegenerate linear transform, which maps D onto A' and $\digamma'$ onto $\digamma$, then A' is an interior point of $\digamma$. Therefore A' is a relative interior point of $\digamma_{A'}$. Then

$$V_{A'} = \{X \in S^{n\times n} | AX = 0, \langle Q', X\rangle = 0\}.$$

To show that $V_{A'} = V$, we need only verify that

$$V' = \{A'^{\frac{1}{2}}XA^{\frac{1}{2}} | X \in S_+^{n\times n}\} = \digamma = \{X \in S_+^{n\times n} | \langle Q', X\rangle = 0\}.$$

For all $F \in \digamma$, since $\digamma = U\digamma' U^{-1}$,

$$F = U \begin{pmatrix} Y' & 0 \\ 0 & 0 \end{pmatrix} U^{-1}$$

for some $r \times r$ positive semidefinite matrix Y'. Let

$$X = U \begin{pmatrix} D'Y'D' & 0 \\ 0 & 0 \end{pmatrix} U^{-1};$$

here $D' = \operatorname{diag}\{\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_r}\}^{-1}, \lambda_i > 0, i = 1, \ldots, r$ are the eigenvalues of A'. Then

$$A'^{\frac{1}{2}}XA'^{\frac{1}{2}} = \left(U\operatorname{diag}\{\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_r}, 0, \ldots, 0\}U^{-1}\right)\left(U\begin{pmatrix} D'Y'D' & 0 \\ 0 & 0 \end{pmatrix}U^{-1}\right),$$

$$\left(U\mathrm{diag}\{\sqrt{\lambda_1},\ldots,\sqrt{\lambda_r},0,\ldots,0\}U^{-1}\right) = U\begin{pmatrix} Y^{'} & 0 \\ 0 & 0 \end{pmatrix}U^{-1} = F.$$

We conclude that $\digamma \subset V'$. It is easy to see that $\dim(\digamma) = \dim(V') = \frac{r(r+1)}{2}$, and we get $\digamma = V'$. Therefore,

$$\{X \in S^{n\times n} | \langle Q', X\rangle = 0\} = Aff(\digamma) = Aff(V') = \{A'^{\frac{1}{2}} X A'^{\frac{1}{2}} | X \in S^{n\times n}\},$$

which implies that

$$V_{A'} = V = \{F''(A')^{-\frac{1}{2}} X | AF''(A')^{-\frac{1}{2}} X = 0, X \in S^{n\times n}\}.$$

Before we prove Theorem 3.3, we introduce some notation of Jordan algebra. Since every symmetric cone K can be realized as a cone of squares in an appropriated Euclidean algebra (see Faraut and Koranyi [9] for details), we can use the Jordan algebra technique to prove Theorem 3.3.

Let V be an Euclidean Jordan algebra and Ω be a cone of invertible squares in V. We define $\langle x, y\rangle = \mathrm{tr}(x \circ y)$ as the canonical scalar product in V. Let $F(x) = -\log\det(x)$, $x \in \Omega$. Then $F''(x) = P(x)^{-1}$; here $F''(x)$ is the Hession of F evaluated at $x \in \Omega$ with respect to the canonical scalar product $\langle , \rangle$. $P(x)$ is the quadratic representation of x. We assume $\mathrm{rank}(V) = r$. When x is on the boundary $\partial\Omega$ of Ω, $\mathrm{rank}(x) = j < r$.

In the following, we fix a Jordan frame $c_1, \ldots, c_r$ and denote $e_j = c_1 + \cdots + c_j$, $V^{(j)} = V(e_j, 1)$. We denote by Ω^j the symmetric cone associated with the subspace $V^{(j)}$, i.e., the interior relative to $V^{(j)}$. Then $\Omega_j \subset \partial\Omega$. The following lemma characterizes the boundary of symmetric cone. For a proof, see Proposition IV.3.1 in Faraut and Koranyi [9].

LEMMA A.1. *For x in $\overline{\Omega}$ the following properties are equivalent:*

(a) *The rank of x is j.*
(b) *$x \in k\Omega^j$ for some k in $K = G \cap O(V)$; here G is the connected component of the identity in $G(\Omega)$ and $G(\Omega)$ denotes the set of automorphisms of Ω.*
(c) *The rank of $P(x)$ is equal to the dimension of $V^{(j)}$.*

Now we assume $x^* \in \overline{\Omega}$ and $\mathrm{rank}(x^*) = j$. From Lemma A.1, we know that $x^* \in k\Omega^j$ for some k in K. It can be verified that $V_{x^*} = kV^{(j)}$. Now we are ready to prove Theorem 3.3.

Proof of Theorem 3.3. From the above analysis, we need only prove $P(x^*)^{\frac{1}{2}} V = kV^{(j)}$. Since $P(x^*)$ is a positive semidefinite linear operator, $P(x^*)^{\frac{1}{2}} V = P(x^*)V$. Therefore we need only prove $P(x^*)V = kV^{(j)}$. From part (b) of Lemma A.1, we know $x^* = k\sum_{i=1}^{j} \lambda_i c_i = kP(a)e_j$, with $a = \sum_{i=1}^{j} \sqrt{\lambda_i} c_i + \sum_{i=j+1}^{r} c_i$. Then $P(x^*) = p(kP(a)e_j) = kP(P(a)e_j)k^* = kP(a)P(e_j)P(a)k^*$; here the second equality follows by Proposition III.5.2 in Faraut and Koranyi [9] and the last equality follows by Proposition II.3.3 in Faraut and Koranyi [9]. Since $P(e_j)$ is the orthogonal projection onto $V^{(j)}$ and $P(a)$ maps $V^{(j)}$ onto $V^{(j)}$, it is easy to see that $P(x^*)V \subset kV^{(j)}$. Since from part (c) of Lemma A.1, we know $\mathrm{rank}(P(x^*)) = \mathrm{rank}(V^{(j)}) = \mathrm{rank}(kV^{(j)})$, we conclude that $P(x^*)V = kV^{(j)}$. We complete the proof. □

Part (a) of Lemma 4.2 holds only because $F''(x^*)^{-\frac{1}{2}} = P(x^*)^{\frac{1}{2}}$ and $P(x^*)$ is the quadratic representation of x^*.

Acknowledgments. The first author would like to thank his advisor, Andrew Sommese, at Notre Dame for his help and support, and Professor Faybusovich for his guidance in the interior-point methods. Professor Yinyu Ye read this paper carefully

and gave some sincere advice; we appreciate his support and encouragement. We would like to thank Mr. Dong Yang for his help in the implementation of our algorithm. We are grateful to the associate editor and the referees for their precious comments.

REFERENCES

[1] P. A. Absil and A. L. Tits, *Newton-KKT interior-point methods for indefinite quadratic programming*, Comput. Optim. Appl., to appear.
[2] F. Alizadeh and D. Goldfarb, *Second-order cone programming*, Math. Program., 95 (2003), pp. 3–51.
[3] A. S. El Bakry, R. A. Tapia, T. Tsuchiya, and Y. Zhang, *On the formulation and theory of Newton interior point methods for nonlinear programming*, J. Optim. Theory Appl., 89 (1996), pp. 507–541.
[4] A. Barvinok, *A Course in Convexity*, AMS, Providence, RI, 2002.
[5] J. F. Bonnans and A. Shapiro, *Perturbation Analysis of Optimization Problems*, Springer-Verlag, New York, 2000.
[6] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, *Trust-Region Methods*, SIAM, Philadelphia, 2000.
[7] A. R. Conn, N. I. M. Gould, D. Orban, and Ph. L. Toint, *A primal-dual trust-region algorithm for nonconvex nonlinear programming*, Math. Program., 87 (2000), pp. 215–249.
[8] I. I. Dikin, *Iterative solution of problems of linear and quadratic programming*, Soviet Math. Dokl., 8 (1967), pp. 674–675.
[9] J. Faraut and A. Koranyi, *Analysis on Symmetric Cones*, Oxford University Press, New York, 1996.
[10] L. Faybusovich and T. Tsuchiya, *Primal-dual algorithms and infinite-dimensional Jordan algebras of finite rank*, Math. Program., 97 (2003), pp. 471–493.
[11] L. Faybusovich and Y. Lu, *Jordan-algebraic aspects of nonconvex optimization over symmetric cone*, Appl. Math. Optim., 53 (2006), pp. 67–77.
[12] A. Forsgren and P. E. Gill, *Primal-dual interior methods for nonconvex nonlinear programming*, SIAM J. Optim., 8 (1998), pp. 1132–1152.
[13] D. M. Gay, M. L. Overton, and M. H. Wright, *A primal-dual interior method for nonconvex nonlinear programming*, in Advances in Nonlinear Programming, Y. Yuan, ed., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 31–56.
[14] N. I. M. Gould, S. Lucidi, M. Roma, and Ph. L. Toint, *Solving the trust-region subproblem using the Lanczos method*, SIAM J. Optim., 9 (1999), pp. 504–525.
[15] N. I. M. Gould, M. E. Hribar, and J. Nocedal, *On the solution of equality constrained quadratic programming problems arising in optimization*, SIAM J. Sci. Comput., 23 (2001), pp. 1376–1395.
[16] J. Griffin and P. E. Gill, *Trust-region interior methods for large-scale optimization*, in Proceedings of the Eighth SIAM Conference on Optimization, Stockholm, Sweden, 2005.
[17] Y. Lu and Y. Yuan, *An Interior-Point Trust-Region Polynomial Algorithm for Convex Programming*, Technical report, Department of Mathematics, University of Notre Dame, Notre Dame, IN, 2006.
[18] Y. E. Nesterov and A. S. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM, Philadelphia, 1994.
[19] F. Rendel and H. Wolkowicz, *A semidefinite framework for trust region subproblem with applications to large scale minimization*, Math. Programming, 77 (1997), pp. 273–299.
[20] J. Renegar, *A Mathematical View of Interior-Point Methods in Convex Programming*, SIAM, Philadelphia, 2001.
[21] M. Rojas, S. A. Santos, and D. C. Sorensen, *A new matrix-free algorithm for the large-scale trust-region subproblem*, SIAM J. Optim., 11 (2000), pp. 611–646.
[22] D. C. Sorensen, *Minimization of a large-scale quadratic function subject to a spherical constraint*, SIAM J. Optim., 7 (1997), pp. 141–161.
[23] T. Steihaug, *The conjugate gradient method and trust regions in large scale optimization*, SIAM J. Numer. Anal., 20 (1983), pp. 626–637.
[24] J. Sun, *A convergence proof for an affine-scaling method for convex quadratic programming without nondegeneracy assumptions*, Math. Programming, 60 (1993), pp. 69–79.
[25] A. L. Tits, A. Wächter, S. Bakhtiari, T. J. Urban, and C. T. Lawrence, *A primal-dual interior-point method for nonlinear programming with strong global and local convergence properties*, SIAM J. Optim., 14 (2003), pp. 173–199.
[26] Ph. L. Toint, *Towards an efficient sparsity exploiting Newton method for minimization*, in

Sparse Matrices and Their Uses, I. S. Duff, ed., Academic Press, London, 1981, pp. 57–88.
[27] R. J. Vanderbei and D. F. Shanno, *An interior-point algorithm for nonconvex nonlinear programming*, Comput. Optim. Appl., 13 (1999), pp. 231–252.
[28] A. Wächter, *An Interior Point Algorithm for Large-Scale Nonlinear Optimization with Applications in Process Engineering*, Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA, 2002.
[29] Y. Ye, *On an affine scaling algorithm for non-convex quadratic programming*, Math. Programming, 52 (1992), pp. 285–300.
[30] Y. Ye, *Interior Point Algorithms: Theory and Analysis*, John Wiley and Sons, New York, 1997.
[31] Y. Ye and S. Zhang, *New results on quadratic minimization*, SIAM J. Optim., 14 (2003), pp. 245–267.
[32] Y. Yuan, *On the truncated conjugate gradient method*, Math. Program., 87 (2000), pp. 561–571.
[33] S. Zhang, *Quadratic maximization and semidefinite relaxation*, Math. Program., 87 (2000), pp. 453–465.

SIAM J. OPTIM.
Vol. 18, No. 1, pp. 87–105

COMPUTING THE STABILITY NUMBER OF A GRAPH VIA LINEAR AND SEMIDEFINITE PROGRAMMING*

JAVIER PEÑA†, JUAN VERA‡, AND LUIS F. ZULUAGA§

Abstract. We study certain linear and semidefinite programming lifting approximation schemes for computing the stability number of a graph. Our work is based on and refines de Klerk and Pasechnik's approach to approximating the stability number via copositive programming [*SIAM J. Optim.*, 12 (2002), pp. 875–892]. We provide a *closed-form* expression for the values computed by the linear programming approximations. We also show that the *exact* value of the stability number $\alpha(G)$ is attained by the semidefinite approximation of order $\alpha(G) - 1$ as long as $\alpha(G) \leq 6$. Our results reveal some sharp differences between the linear and the semidefinite approximations. For instance, the value of the linear programming approximation of any order is strictly larger than $\alpha(G)$ whenever $\alpha(G) > 1$.

Key words. stability number, copositivity, polynomials, lifting procedures

AMS subject classifications. 90C35, 90C22, 90C05

DOI. 10.1137/05064401X

1. Introduction.

The maximum stable set problem is a central problem in combinatorial optimization and has been the subject of extensive study. The survey by Bomze et al. [2] gives an overview of a variety of approaches to the maximum clique problem, which is equivalent to the maximum stable set problem. In addition to being a classical NP-hard problem, the maximum stable set is among the provably hardest combinatorial problems to approximate [7, 8].

The maximum stable set problem has a straightforward 0–1 integer programming formulation and hence is a natural candidate for application of integer programming. Indeed, lift-and-project approaches provide interesting insight into this problem. For instance, the lift-and-project procedures of Balas, Ceria, and Cornuéjols [1] and of Lovász and Schrijver [12] give relaxations of the stable set polytope that already satisfy a number of valid inequalities. Furthermore, the repeated lift-and-project procedures of Balas, Ceria, and Cornuéjols and of Lovász and Schrijver yield a *finite* sequence of increasingly tighter relaxations of the stable set polytope. Indeed, for a graph G with stability number $\alpha(G)$, the Lovász and Schrijver linear lifting procedure yields the stable set polytope if it is applied $n - \alpha(G) - 1$ times. The stronger Lovász and Schrijver semidefinite lifting procedure yields the stable set polytope when applied $\alpha(G)$ times.

Using a different mathematical programming formulation of the stable set problem, de Klerk and Pasechnik [4] proposed two alternative approximation schemes to $\alpha(G)$. Their approach is based on a characterization of $\alpha(G)$ via copositive programming (cf. (2)) combined with a successive approximation procedure for the copositive

*Received by the editors November 1, 2005; accepted for publication (in revised form) October 12, 2006; published electronically February 2, 2007.
http://www.siam.org/journals/siopt/18-1/64401.html

†Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA 15213-3890 (jfp@andrew.cmu.edu). This author was supported by NSF grant CCF-0092655.

‡Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213 (jvera@andrew.cmu.edu). This author was partially supported by NSF grant CCF-0092655.

§Faculty of Business Administration, University of New Brunswick, PO Box 4400, Fredericton, NB E38 5X9, Canada (lzuluaga@unb.ca). This author was supported by NSERC grant 290377 and FDF, University of New Brunswick.

cone inspired by the work of Parrilo [14]. More specifically, de Klerk and Pasechnik [4] define two sequences of cones $\mathcal{K}_n^0 \subseteq \mathcal{K}_n^1 \subseteq \cdots$ and $\mathcal{C}_n^0 \subseteq \mathcal{C}_n^1 \subseteq \cdots$ that converge to the n-dimensional copositive cone $\mathcal{C}_n$ and define $\vartheta^{(r)}(G)$ and $\zeta^{(r)}(G)$ as the upper bounds on $\alpha(G)$ obtained by replacing the copositive cone $\mathcal{C}_n$ by $\mathcal{K}_n^r$ and $\mathcal{C}_n^r$, respectively. Each $\vartheta^{(r)}(G)$ can be computed via semidefinite programming, and each $\zeta^{(r)}(G)$ can be computed via linear programming. De Klerk and Pasechnik establish several properties of the approximations $\vartheta^{(r)}(G)$ and $\zeta^{(r)}(G)$. They show that $\vartheta^{(0)}(\cdot)$ coincides with Schrijver's ϑ'-function. They also show that $\vartheta^{(1)}(\cdot) = \alpha(\cdot)$ for odd cycles and complements of triangle-free graphs. In addition, they prove that the approximations $\zeta^{(r)}(G)$ satisfy $\lfloor \zeta^{(r)}(G) \rfloor = \alpha(G)$ as long as $r \geq \alpha(G)^2$ and $\zeta^{(r)}(G) < \infty$ only if $r \geq \alpha(G) - 1$.

We refine the results of de Klerk and Pasechnik concerning the approximations $\zeta^{(r)}(G)$ and $\vartheta^{(r)}(G)$ to the stability number $\alpha(G)$. We derive a closed-form expression for $\zeta^{(r)}(G)$ in terms of r and $\alpha(G)$ (Theorem 1). The closed-form expression readily yields several interesting properties of $\zeta^{(r)}(G)$ (Corollaries 2, 3, and 4).

We also define a third sequence of cones $\mathcal{Q}_n^0 \subseteq \mathcal{Q}_n^1 \subseteq \cdots$ that converges to $\mathcal{C}_n$ with $\mathcal{C}_n^r \subseteq \mathcal{Q}_n^r \subseteq \mathcal{K}_n^r$, $r = 0, 1, \ldots$. This in turn yields a third sequence of approximations $\nu^{(r)}(G)$ to $\alpha(G)$ with $\vartheta^{(r)}(G) \leq \nu^{(r)}(G) \leq \zeta^{(r)}(G)$. Like $\vartheta^{(r)}(G)$, each $\nu^{(r)}(G)$ can be computed via semidefinite programming. The approximations $\nu^{(r)}(\cdot)$ satisfy an interesting recursive inequality (Theorem 6). Such inequality implies $\nu^{(r)}(G) = \alpha(G)$ as long as $r \geq \alpha(G) - 1$ for $\alpha(G) \leq 6$ (Corollary 7). In particular, the latter gives a partial solution to a conjecture of de Klerk and Pasechnik [4, Conjecture 5.1].

Our results reveal some sharp differences between the two approximation schemes $\zeta^{(r)}(G)$ and $\nu^{(r)}(G)$. For instance, $\zeta^{(r)}(G) \downarrow \alpha(G)$, but $\zeta^{(r)}(G) > \alpha(G)$ for all r whenever $\alpha(G) > 1$ (Corollary 2). Furthermore, as it was previously established by de Klerk and Pasechnik, $\zeta^{(r)}(G) = \infty$ for $r < \alpha(G) - 1$ (Corollary 3). By contrast, $\nu^{(r)}(G) \leq \chi(\overline{G}) \leq n$ for all r (Proposition 14), where $\chi(\overline{G})$ denotes the chromatic number of the complement of G.

It is interesting to note that our results guarantee the convergence of $\zeta^{(r)}(G)$ and $\nu^{(r)}(G)$ to $\alpha(G)$ *without* relying on the convergence of $\mathcal{Q}_n^r$ and $\mathcal{C}_n^r$ to the copositive cone $\mathcal{C}_n$ (Corollary 2).

The paper is organized as follows. In section 2 we introduce the basic terminology and notation. In section 3 we present the closed-form expression for $\zeta^{(r)}(\cdot)$ and discuss some of its consequences. In section 4 we discuss the successive approximations $\nu^{(r)}(\cdot)$. Finally in section 5 we specialize some of our results to three special classes of graphs, namely, the graphs whose stability number coincides with the chromatic number of their complement, the cycles, and the complements of cycles.

During the completion of this paper, we learned about the related independent work by Gvozdenović and Laurent [6]. Gvozdenović and Laurent studied some properties of the approximations $\vartheta^{(r)}(G)$. In particular, they establish a connection with the approximations obtained by applying Lasserre's lift-and-project procedure [9, 10] and also give a partial solution to [4, Conjecture 5.1] that is slightly stronger than ours.

2. Preliminaries. Throughout the paper $G = (V, E)$ will denote a loopless undirected graph with vertex set $V = \{1, \ldots, n\}$. A subset $S \subseteq V$ is *stable* if $\{i, j\} \notin E$ for all $i, j \in S$. The *stability number* $\alpha(G)$ is the cardinality of a stable set of maximum size in G.

Let $\mathbb{S}^n$ denote the space of symmetric $n \times n$ matrices. The *positive semidefinite cone* $\mathbb{S}_+^n \subseteq \mathbb{S}^n$ is

$$\mathbb{S}_+^n := \{X \in \mathbb{S}^n : u^{\mathrm{T}} X u \geq 0 \text{ for all } u \in \mathbb{R}^n\}.$$

Following the usual convention in the semidefinite programming literature, we will write $X \succeq 0$ as shorthand for $X \in \mathbb{S}^n_+$. In addition, we will write $X \geq 0$ to indicate that every entry of X is nonnegative.

The *copositive cone* $\mathcal{C}_n \subseteq \mathbb{S}^n$ is

$$\mathcal{C}_n := \{X \in \mathbb{S}^n : u^{\mathrm{T}} X u \geq 0 \ \text{ for all } \ u \in \mathbb{R}^n_+\}.$$

Throughout the paper $A(G) \in \mathbb{S}^n$ will denote the *adjacency* matrix of the graph G; i.e., the (i,j) entry of $A(G)$ is 1 if $\{i,j\} \in E$ and it is 0 otherwise. Also, e will denote the vector of all ones, and I will denote the identity matrix, whose dimensions will be clear from the context.

Our work relies on the following inequality, which can be easily verified:

$$\alpha(G) \leq \inf\{\lambda : \lambda(I + A(G)) - ee^{\mathrm{T}} \in \mathcal{C}_n\}. \tag{1}$$

De Klerk and Pasechnik showed that indeed the following stronger identity holds [4, Corollary 2.4]:

$$\alpha(G) = \min\{\lambda : \lambda(I + A(G)) - ee^{\mathrm{T}} \in \mathcal{C}_n\}. \tag{2}$$

It is easy to see that $\{P + N : P, N \in \mathbb{S}^n,\ P \succeq 0,\ N \geq 0\} \subseteq \mathcal{C}_n$. Furthermore, it is known [5] that this inclusion is strict for $n \geq 5$. There is not a simple description of the copositive cone, and in fact the problem of deciding if a matrix is copositive is known to be NP-hard [13]. A way of addressing this difficulty was proposed by Parrilo [14]. For a given symmetric matrix $M \in \mathbb{S}^n$, consider the four degree form (homogeneous polynomial)

$$P_M(x) := \sum_{i=1}^{n} \sum_{j=1}^{n} M_{ij} x_i^2 x_j^2.$$

Observe that M is copositive if and only if $P_M(x) \geq 0$ for all $x \in \mathbb{R}^n$.

Parrilo proposed the following *sum-of-squares* (sos) approximation scheme for the copositive cone: Define the cone $\mathcal{K}^r_n \subseteq \mathcal{C}_n$ as

$$\mathcal{K}^r_n := \left\{ M \in \mathbb{S}^n : \left(\sum_{i=1}^{n} x_i^2 \right)^r P_M(x) \text{ is a sos} \right\}. \tag{3}$$

It is easy to see that the sos condition can be recast in terms of linear matrix inequalities (LMI). Therefore, membership in each $\mathcal{K}^r_n$ can be phrased in terms of LMI. In particular, Parrilo showed that $M \in \mathcal{K}^0_n$ if and only if $M = P + N$ for $P \succeq 0$, $N \geq 0$. Parrilo also showed that $M \in \mathcal{K}^1_n$ if the following system of LMI has a solution:

$$\begin{aligned} M - \Lambda^i &\succeq 0, && i = 1, \ldots, n, \\ \Lambda^i_{ii} &= 0, && i = 1, \ldots, n, \\ \Lambda^i_{jj} + \Lambda^j_{ji} + \Lambda^j_{ij} &= 0, && i \neq j, \\ \Lambda^i_{jk} + \Lambda^j_{ik} + \Lambda^k_{ij} &\geq 0, && i, j, k \ \text{ all different.} \end{aligned} \tag{4}$$

Bomze and de Klerk [3] showed that indeed $M \in \mathcal{K}^1_n$ if and only if (4) has a solution.

De Klerk and Pasechnik define

$$\vartheta^{(r)}(G) := \min\{\lambda : \lambda(I + A(G)) - ee^{\mathrm{T}} \in \mathcal{K}^r_n\}.$$

Since each $\mathcal{K}_n^r$ can be defined in terms of LMI, the approximation $\vartheta^{(r)}(G)$ can be computed via semidefinite programming.

De Klerk and Pasechnik also define the cone $\mathcal{C}_n^r \subseteq \mathcal{C}_n$ as

$$\mathcal{C}_n^r := \left\{ P \in \mathbb{S}^n : \left(\sum_{i=1}^n x_i\right)^r x^{\mathrm{T}} P x \text{ has nonnegative coefficients} \right\},$$

and

$$\zeta^{(r)}(G) := \min\{\lambda : \lambda(I + A(G)) - ee^{\mathrm{T}} \in \mathcal{C}_n^r\}.$$

It is easy to see that $\mathcal{C}_n^r \subseteq \mathcal{K}_n^r \subseteq \mathcal{C}_n$ so $\alpha(G) \leq \vartheta^{(r)}(G) \leq \zeta^{(r)}(G)$. Furthermore, since each $\mathcal{C}_n^r$ is polyhedral, the approximation $\zeta^{(r)}(G)$ can be computed via linear programming. De Klerk and Pasechnik [4] established the following interesting properties of $\vartheta^{(r)}(\cdot)$ and $\zeta^{(r)}(\cdot)$: The function $\vartheta^{(0)}(\cdot)$ coincides with Schrijver's ϑ'-function, and $\vartheta^{(1)}(\cdot) = \alpha(\cdot)$ for odd cycles and complements of triangle-free graphs. They also proved that $\lfloor \zeta^{(r)}(G) \rfloor = \alpha(G)$ for $r \geq \alpha(G)^2$ and $\zeta^{(r)}(G) < \infty$ only if $r \geq \alpha(G) - 1$.

3. Linear programming approximations. From the above definition of the linear approximation $\zeta^{(r)}(G)$, the fact that $\mathcal{C}_n^0 \subseteq \mathcal{C}_n^1 \subseteq \cdots \subseteq \mathcal{C}_n$, and (1), it follows that

$$\zeta^{(0)}(G) \geq \zeta^{(1)}(G) \geq \cdots \geq \alpha(G).$$

Theorem 1 below gives a closed-form expression for $\zeta^{(r)}(G)$ in terms of r and $\alpha(G)$.

Throughout this section, the binomial coefficient $\binom{a}{2}$ is to be understood as $\frac{a(a-1)}{2}$ for any nonnegative integer a. In particular, $\binom{a}{2} = 0$ for $a = 0$ and $a = 1$. Also, by convention $a/0 = +\infty$ for $a > 0$.

THEOREM 1. *Assume $r + 2 = u\alpha(G) + v$, where u, v are nonnegative integers with $v < \alpha(G)$. Then*

$$\zeta^{(r)}(G) = \frac{\binom{r+2}{2}}{\binom{u}{2}\alpha(G) + vu}.$$

COROLLARY 2. *$\zeta^{(r)}(G) \downarrow \alpha(G)$. Furthermore, if $\alpha(G) > 1$, then $\zeta^{(r)}(G) > \alpha(G)$ for all nonnegative integers r.*

Proof. To simplify notation write α as a shorthand for $\alpha(G)$. Let u, v be nonnegative integers such that $r + 2 = u\alpha + v$ and $v < \alpha$. Then for $\epsilon \geq 0$

$$(\alpha + \epsilon)\left(\binom{u}{2}\alpha + vu\right) - \binom{r+2}{2} = \frac{u\alpha(1 + \epsilon u - \epsilon - \alpha) + v(2\epsilon u - v + 1)}{2}. \tag{5}$$

Let $\epsilon > 0$ be given. Then from (5) it follows that for r sufficiently large

$$(\alpha + \epsilon)\left(\binom{u}{2}\alpha + vu\right) > \binom{r+2}{2}.$$

Thus Theorem 1 yields $\limsup_{r\to\infty} \zeta^{(r)}(G) \leq \alpha(G)$. Since $\zeta^{(r)}(G) \geq \alpha(G)$ for all r, then indeed $\lim_{r\to\infty} \zeta^{(r)}(G) = \alpha(G)$.

For the second part, just observe that for $\epsilon = 0$ and $\alpha > 1$ the right-hand side of (5) is negative. Thus again Theorem 1 yields $\zeta^{(r)}(G) > \alpha(G)$ whenever $\alpha(G) > 1$. ∎

From Theorem 1, we can also recover two key properties of $\zeta^{(r)}(G)$ due to de Klerk and Pasechnik [4]. The following two corollaries are slight refinements of [4, Theorem 4.2] and [4, Theorem 4.1], respectively.

COROLLARY 3. *$\zeta^{(r)}(G) < \infty$ if and only if $r \geq \alpha(G) - 1$.*

Proof. By Theorem 1, $\zeta^{(r)}(G) = \infty$ if and only if $\binom{u}{2}\alpha(G) + vu = 0$. The latter occurs if and only if $u = 0$ or $u = 1$ and $v = 0$, i.e., if and only if $r + 2 \leq \alpha(G)$. ∎

COROLLARY 4. *$\lfloor \zeta^{(r)}(G) \rfloor = \alpha(G)$ if and only if $r \geq \alpha(G)^2 - 1$.*

Proof. To simplify notation write α as a shorthand for $\alpha(G)$. Since $\zeta^{(r)}(G) \geq \alpha$, it follows that $\lfloor \zeta^{(r)}(G) \rfloor = \alpha$ if and only if $\zeta^{(r)}(G) < \alpha + 1$. Since $\zeta^{(r)}(G)$ is nonincreasing in r, it suffices to show that

$$\zeta^{(\alpha^2-2)}(G) = \alpha + 1 \quad \text{and} \quad \zeta^{(\alpha^2-1)}(G) < \alpha + 1.$$

But this readily follows from Theorem 1. ∎

The proof of Theorem 1 relies on Lemma 5, which gives a bound on the number of edges of any induced subgraph of $G = (V, E)$. Given a graph G and $S \subseteq V$, let $E_G[S]$ be the set of edges of G with both endpoints in S.

LEMMA 5. *Let $G = (V, E)$ be a graph, and assume $S \subseteq V$. Let u, v be nonnegative integers such that $|S| = u\alpha(G) + v$ and $v < \alpha(G)$. Then $|E_G[S]| \geq \binom{u}{2}\alpha(G) + vu$.*

Proof. Proceed by induction on u. The result is trivial for $u = 0$. Thus assume $|S| \geq \alpha(G)$. Let I be a maximal stable subset of S, and let $S_1 = S \setminus I$. Since I is stable, $|I| \leq \alpha(G)$ and $|S_1| \geq |S| - \alpha(G) = (u-1)\alpha(G) + v$. Furthermore, since I is maximal in S, every vertex in S_1 is connected to some vertex in I. Therefore there are at least $|S_1|$ edges between S_1 and I. Then

$$|E_G[S]| \geq |S_1| + |E_G[S_1]| \geq |S| - \alpha(G) + |E_G[S_1]|.$$

But by induction hypothesis $|E_G[S_1]| \geq \binom{u-1}{2}\alpha(G) + v(u-1)$. Therefore

$$|E_G[S]| \geq |S| - \alpha(G) + \binom{u-1}{2}\alpha(G) + v(u-1) = \binom{u}{2}\alpha(G) + vu. \qquad \blacksquare$$

In the proof below we use the following convenient notation: For a positive integer d, we write $[d]$ as shorthand for the set $\{1, \ldots, d\}$. Also, K_r denotes the complete graph with r vertices.

Proof of Theorem 1. Let $B := I + A(G)$. For any $1 \leq i < j \leq r + 2$ we have

$$\left(\sum_{k=1}^{n} x_k\right)^r x^{\mathrm{T}}(\lambda B - ee^{\mathrm{T}})x = \sum_{s \in [n]^{r+2}} (\lambda B_{s_i,s_j} - 1)x_{s_1} \ldots x_{s_{r+2}}.$$

Thus, adding over all possible pairs (i, j) with $1 \leq i < j \leq r + 2$, we get

$$\begin{aligned}\left(\sum_{k=1}^{n} x_k\right)^r x^{\mathrm{T}}(\lambda B - ee^{\mathrm{T}})x &= \sum_{s \in [n]^{r+2}} \left(\frac{\lambda}{\binom{r+2}{2}} \sum_{1 \leq i < j \leq r+2} B_{s_i,s_j} - 1\right) x_{s_1} \ldots x_{s_{r+2}} \\ &= \sum_{1 \leq s_1 \leq \cdots \leq s_{r+2} \leq n} perm(s) \left(\frac{\lambda}{\binom{r+2}{2}} \sum_{1 \leq i < j \leq r+2} B_{s_i,s_j} - 1\right) x_{s_1} \ldots x_{s_{r+2}},\end{aligned}$$

where $perm(s)$ is the number of different permutations of $s = (s_1, \ldots, s_{r+2})$.

Therefore

$$\zeta^{(r)}(G) \leq \lambda \Leftrightarrow \lambda \sum_{1\leq i<j\leq r+2} B_{s_i,s_j} \geq \binom{r+2}{2} \text{ for all } s \in [n]^{r+2}. \tag{6}$$

Fix $s \in [n]^{r+2}$. Let $H = K_{r+2} \times G$, i.e., the strong product of the graphs K_{r+2} and G. Let $S = \{(i, s_i) : i = 1, \ldots, r+2\} \subseteq V(H)$. Notice that

$$B_{s_i,s_j} = 1 \Leftrightarrow \{s_i, s_j\} \in E(G) \text{ or } s_i = s_j \Leftrightarrow \{(i, s_i), (j, s_j)\} \in E(H).$$

Hence $\sum_{1\leq i<j\leq r+2} B_{s_i,s_j} = |E_H[S]|$. Since $\alpha(H) = \alpha(G)$, Lemma 5 yields

$$\sum_{1\leq i<j\leq r+2} B_{s_i,s_j} \geq \binom{u}{2}\alpha(G) + vu.$$

This can be done for each $s \in [n]^{r+2}$. Hence it follows from (6) that

$$\zeta^{(r)}(G) \leq \frac{\binom{r+2}{2}}{\binom{u}{2}\alpha(G) + vu}.$$

On the other hand, assume $\{v_1, \ldots, v_{\alpha(G)}\}$ is a maximal stable set in G. Define $s \in [n]^{r+2}$ by putting $s_{k\alpha(G)+l} = v_l$ for $k = 0, 1, \ldots, u-1$ and $l = 1, \ldots, \alpha(G)$ and for $k = u$ and $l = 1, \ldots, v$. Then

$$B_{s_i,s_j} = 1 \Leftrightarrow i \equiv j \mod (\alpha(G)),$$

and consequently

$$\sum_{1\leq i<j\leq r+2} B_{s_i,s_j} = \binom{u+1}{2}v + \binom{u}{2}(\alpha(G) - v) = \binom{u}{2}\alpha(G) + vu.$$

Therefore, again from (6) it follows that $\zeta^{(r)}(G) \geq \binom{r+2}{2}/\left(\binom{u}{2}\alpha(G) + vu\right)$. ∎

4. Semidefinite programming approximations. Define the sequences of cones $\mathcal{E}_n^r \subseteq \mathbb{R}[x_1, \ldots, x_n]$, $\mathcal{Q}_n^r \subseteq \mathbb{S}^n$, $r = 0, 1, \ldots$, as follows. For $r = 0, 1, \ldots$ let

$$\mathcal{E}_n^r := \left\{ \sum_{\beta\in\mathbb{N}^n,\, |\beta|=r} x^\beta x^{\mathrm{T}}(P_\beta + N_\beta)x : P_\beta, N_\beta \in \mathbb{S}^n,\, P_\beta \succeq 0,\, N_\beta \geq 0 \right\}, \tag{7}$$

where for a given $\beta \in \mathbb{N}^n$, $|\beta| := \beta_1 + \cdots + \beta_n$ and $x^\beta := x_1^{\beta_1} \ldots x_n^{\beta_n}$. Also let

$$\mathcal{Q}_n^r := \left\{ B \in \mathbb{S}^n : \left(\sum_{i=1}^n x_i\right)^r x^{\mathrm{T}} B x \in \mathcal{E}_n^r \right\}. \tag{8}$$

Notice that $\mathcal{E}_n^r$ contains all homogeneous polynomials of degree $r+2$ with nonnegative coefficients. Therefore, by (8), we have $\mathcal{C}_n^r \subseteq \mathcal{Q}_n^r$, $r = 0, 1, 2, \ldots$.

4.1. LMI description of $\mathcal{Q}_n^r$. From (7) and (8) it follows that membership in $\mathcal{Q}_n^r$ can be written in terms of an LMI in the matrix variables $P_\beta, N_\beta \in \mathbb{S}^n$, $P_\beta \succeq 0$, $N_\beta \geq 0$. This involves $2\binom{n+r-1}{r} = 2|\{\beta \in \mathbb{N}^n,\ |\beta| = r\}|$ matrices of size $n \times n$. More precisely, $M \in \mathcal{Q}_n^r$ if and only if there exist $P_\beta, N_\beta \in \mathbb{S}^n$, $P_\beta \succeq 0$, $N_\beta \geq 0$ such that

$$\sum_{\beta \in \mathbb{N}^n,\, |\beta| = r} x^\beta x^{\mathrm{T}} M x = \sum_{\beta \in \mathbb{N}^n,\, |\beta| = r} x^\beta x^{\mathrm{T}} (P_\beta + N_\beta) x. \tag{9}$$

Notice that the equality in (9) is equivalent to the equality of the coefficients of the two polynomials appearing in each side of the expression. Since each coefficient is a linear combination of the entries of the involved matrices, the equality in (9) is a linear system of equations in the entries of matrix M and the matrices P_β, N_β.

Notice that (9) can be written as

$$\sum_{\beta \in \mathbb{N}^n,\, |\beta| = r} x^\beta x^{\mathrm{T}} M x \preceq \sum_{\beta \in \mathbb{N}^n,\, |\beta| = r} x^\beta x^{\mathrm{T}} P_\beta x \tag{10}$$

involving only the $\binom{n+r-1}{r}$ matrix variables $P_\beta \in \mathbb{S}_+^n$. Here we use the sign "$\preceq$" to indicate that the vector of coefficients of the polynomial in the left-hand side is componentwise less than or equal to the vector of coefficients of the polynomial in the right-hand side. Furthermore, by grouping identical monomials, (10) can be written in the following slightly more concise form involving fewer matrix variables $P_\beta \in \mathbb{S}_+^n$:

$$\sum_{\beta \in \mathbb{N}^n,\, |\beta| = r} x^\beta x^{\mathrm{T}} M x \preceq \sum_{\substack{\beta \in \mathbb{N}^n,\, |\beta| = r \\ \beta_1 \leq \beta_2 \leq \cdots \leq \beta_r}} x^\beta x^{\mathrm{T}} P_\beta x. \tag{11}$$

The condition (11) is a linear system of inequalities in the entries of M and P_β. It states that the vector of coefficients in the left-hand side, which is a linear combination of entries of M, should be componentwise less than or equal to the vector of coefficients in the right-hand side, which is a linear combination of the entries of the matrices P_β.

It is insightful to compare the cones $\mathcal{Q}_n^r$ and $\mathcal{K}_n^r$. To that end, we note that by [17, Proposition 9] a homogeneous polynomial $p(x)$ of degree $r+2$ satisfies the condition

$$p(x_1^2, \ldots, x_n^2) \text{ is a sos}$$

if and only if

$$p(x) \in \mathcal{F}_n^r := \left\{ \sum_{\beta \in \mathbb{N}^n, |\beta| \leq r+2} x^\beta q_\beta(x) : q_\beta(x) \text{ is a sos} \right\}. \tag{12}$$

Hence the set $\mathcal{K}_n^r$ defined by (3) can also be described as

$$\mathcal{K}_n^r = \left\{ B \in \mathbb{S}^n : \left(\sum_{i=1}^n x_i \right)^r x^{\mathrm{T}} B x \in \mathcal{F}_n^r \right\}. \tag{13}$$

From (7) and (12) it follows that $\mathcal{E}_n^0 = \mathcal{F}_n^0$, $\mathcal{E}_n^1 = \mathcal{F}_n^1$, and $\mathcal{E}_n^r \subseteq \mathcal{F}_n^r$, $r = 2, \ldots$. Therefore from (8) and (13) we get $\mathcal{Q}_n^0 = \mathcal{K}_n^0$, $\mathcal{Q}_n^1 = \mathcal{K}_r^1$, and $\mathcal{Q}_n^r \subseteq \mathcal{K}_n^r$, $r = 2, \ldots$. In

addition, via a standard limiting argument it follows that each cone $\mathcal{Q}_n^r$ is closed. It should be noted that the identities $\mathcal{E}_n^0 = \mathcal{F}_n^0$, $\mathcal{E}_n^1 = \mathcal{F}_n^1$ were first derived (in a slightly different form) by Parrilo [14].

As it is now well documented (see, e.g., [9, 14, 17]), the sos condition can be written as an LMI. Specifically, by letting $x^{[d]}$ denote the vector of monomials of degree d, it follows that a $2d$-degree homogeneous polynomial $q(x)$ is sos if and only if there exists $P \in \mathbb{S}_+^{\binom{n+d-1}{d}}$ such that

$$q(x) = (x^{[d]})^{\mathrm{T}} P x^{[d]}.$$

It thus follows that membership in $\mathcal{K}_n^r$ can be written as an LMI. More precisely, from (12) it follows that $M \in \mathcal{K}_n^r$ if and only if for $k = 0, 1, \ldots, \lfloor \frac{r+2}{2} \rfloor$ there exist $P_{k,\beta} \in \mathbb{S}_+^{\binom{n+k-1}{k}}$ such that

$$\sum_{\beta \in \mathbb{N}^n,\, |\beta| = r} x^\beta x^{\mathrm{T}} M x = \sum_{k=0}^{\lfloor \frac{r+2}{2} \rfloor} \sum_{\beta \in \mathbb{N}^n,\, |\beta| = r+2-2k} x^\beta (x^{[k]})^{\mathrm{T}} P_{k,\beta} x^{[k]}. \tag{14}$$

Again by grouping identical monomials, (14) can be written in the following slightly more concise form:

$$\sum_{\beta \in \mathbb{N}^n,\, |\beta| = r} x^\beta x^{\mathrm{T}} M x \preceq \sum_{k=1}^{\lfloor \frac{r+2}{2} \rfloor} \sum_{\substack{\beta \in \mathbb{N}^n,\, |\beta| = r+2-2k \\ \beta_1 \le \beta_2 \le \cdots \le \beta_{r+2-2k}}} x^\beta (x^{[k]})^{\mathrm{T}} P_{k,\beta} x^{[k]}. \tag{15}$$

For $r \geq 2$ the LMI description (11) involves a substantially lower number of matrix variables than (15). To illustrate the difference, consider the case $r = 4$. We have $M \in \mathcal{Q}_n^4$ if and only if there exist $P_{ijk\ell} \in \mathbb{S}_+^n$ such that

$$\left(\sum_{i=1}^n x_i \right)^4 x^{\mathrm{T}} M x \preceq \sum_{1 \le i \le j \le k \le \ell \le n} x_i x_j x_k x_\ell \, x^{\mathrm{T}} P_{ijk\ell} x.$$

On the other hand, $M \in \mathcal{K}_n^4$ if and only if there exist $P_{ijk\ell} \in \mathbb{S}_+^n$, $P_{ij} \in \mathbb{S}_+^{\binom{n+1}{2}}$, and $P \in \mathbb{S}_+^{\binom{n+2}{3}}$ such that

$$\left(\sum_{i=1}^n x_i \right)^4 x^{\mathrm{T}} M x \preceq \sum_{1 \le i \le j \le k \le \ell \le n} x_i x_j x_k x_\ell \, x^{\mathrm{T}} P_{ijk\ell} x + \sum_{1 \le i \le j \le n} x_i x_j (x^{[2]})^{\mathrm{T}} P_{ij} x^{[2]} + (x^{[3]})^{\mathrm{T}} P x^{[3]}.$$

4.2. New semidefinite programming approximations to $\alpha(G)$. In analogy to $\zeta^{(r)}(G)$ and $\vartheta^{(r)}(G)$ we define

$$\nu^{(r)}(G) := \min\{\lambda : \lambda(I + A(G)) - ee^{\mathrm{T}} \in \mathcal{Q}_n^r\}. \tag{16}$$

We note that the above minimum is indeed always attained. This follows by applying the same conic duality argument used in [4, section 4]: Both (16) and its dual

are strictly feasible. Notice also that as $\vartheta^{(r)}(G)$, the approximation $\nu^{(r)}(G)$ can be computed via semidefinite programming. The semidefinite program that computes $\nu^{(r)}(G)$ involves only $n \times n$ matrices because the LMI description (11) of $\mathcal{Q}_n^r$ uses $n \times n$ matrices only. By contrast, the LMI description (15) of $\mathcal{K}_n^r$ uses the same number of $n \times n$ matrices in addition to some matrices of size $\binom{n+1}{2} \times \binom{n+1}{2}$, some matrices of size $\binom{n+2}{3} \times \binom{n+2}{3}$, and so on. Henceforth, the semidefinite program that computes $\nu^{(r)}(G)$ is simpler than the one that computes $\vartheta^{(r)}(G)$. Some `matlab` code that constructs the relevant semidefinite programs for $\nu^{(r)}(G)$ and $\vartheta^{(r)}(G)$ in `SeDuMi` format [16] is available at http://www.andrew.cmu.edu/user/jfp/alpha.html. We have used this code for the numerical experiments discussed in section 5.

Observe that $\nu^{(r)}(G) \geq \vartheta^{(r)}(G)$ because $\mathcal{Q}_n^r \subseteq \mathcal{K}_n^r$. Furthermore, the examples in section 5 show that for $r \geq 2$ there are graphs G such that $\nu^{(r)}(G) > \vartheta^{(r)}(G)$.

Since $\mathcal{Q}_n^0 \subseteq \mathcal{Q}_n^1 \subseteq \cdots \subseteq \mathcal{C}_n$ and $\mathcal{C}_n^r \subseteq \mathcal{Q}_n^r$, Corollary 2 implies that

$$\nu^{(r)}(G) \downarrow \alpha(G).$$

Corollary 7 below shows that indeed $\nu^{(r)}(G) = \alpha(G)$ for $r \geq \alpha(G) - 1$ as long as $\alpha(G) \leq 6$.

Given $v \in \{1, \ldots, n\}$ let $v^\perp$ be the union of the neighborhood of v with itself, i.e., $v^\perp := \{j : \{v, j\} \in E\} \cup \{v\}$. Given a set $S \subseteq V$ let $S^\perp := \bigcup_{v \in S} v^\perp$. Let $G \setminus S^\perp$ denote the induced subgraph of G with vertex set $V \setminus S^\perp$, i.e., the graph with vertex set $V \setminus S^\perp$ and edge set $\{\{i, j\} \in E : i, j \in V \setminus S^\perp\}$.

Observe that $\alpha(G)$ satisfies the following relationship as long as $\alpha(G) > r$:

$$\alpha(G) = r + \max_{S \subseteq V \text{ stable}, |S| = r} \alpha(G \setminus S^\perp).$$

The approximations $\nu^{(r)}(\cdot)$ in turn satisfy the related inequality in Theorem 6 below. We note that Theorem 6 generalizes [4, Theorem 5.3].

THEOREM 6. *For $r = 1, 2, 3$ and $\alpha(G) > r$,*

$$\nu^{(r)}(G) \leq r + \max_{S \subseteq V \text{ stable}, |S| = r} \nu^{(0)}(G \setminus S^\perp).$$

Furthermore, the above inequality also holds for $r = 4, 5$ if $\alpha(G) \leq 6$.

COROLLARY 7. *If $r \geq \alpha(G) - 1$ and $\alpha(G) \leq 6$, then $\nu^{(r)}(G) = \alpha(G)$.*

Proof. Since $\nu^{(r)}(G) \downarrow \alpha(G)$ it suffices to show that $\nu^{(\alpha(G)-1)}(G) \leq \alpha(G)$ as long as $\alpha(G) \leq 6$. By Theorem 6 applied to $r = \alpha(G) - 1$ we have

$$\nu^{(\alpha(G)-1)}(G) \leq \alpha(G) - 1 + \max_{S \subseteq V \text{ stable}, |S| = \alpha(G)-1} \nu^{(0)}(G \setminus S^\perp). \tag{17}$$

Notice that $\nu^{(0)}(K) = 1 = \alpha(K)$ for every complete graph K. Notice also that, for each stable set $S \subseteq V$ with $|S| = \alpha(G) - 1$, the subgraph $G \setminus S^\perp$ is a complete graph, and thus $\nu^{(0)}(G \setminus S^\perp) = 1$. Therefore (17) yields

$$\nu^{(\alpha(G)-1)}(G) \leq \alpha(G) - 1 + 1 = \alpha(G). \qquad \square$$

Since $\mathcal{Q}_n^r \subseteq \mathcal{K}_n^r \subseteq \mathcal{C}_n$, we have $\alpha(G) \leq \vartheta^{(r)}(G) \leq \nu^{(r)}(G)$. Thus Corollary 7 yields $\vartheta^{(\alpha(G)-1)}(G) = \alpha(G)$ for $\alpha(G) \leq 6$. This gives a partial solution to [4, Conjecture 5.1].

The proof technique of Theorem 6 also yields the following interesting result.

THEOREM 8. *Let $u \in V$ be such that $u^\perp$ induces a clique in G. Then for $r = 1, 2, 3$ and $\alpha(G) > r$,*

$$\nu^{(r-1)}(G) \leq r + \max_{S \subseteq V \text{ stable}, |S|=r,\, u \in S} \nu^{(0)}(G \setminus S^\perp).$$

Furthermore, the above inequality also holds for $r = 4, 5$ if $\alpha(G) \leq 6$.

Theorems 6 and 8 follow from Lemmas 10 and 11 below, which are interesting on their own. Before stating these lemmas, we introduce some convenient notation: We will write B as shorthand for $I + A(G)$. For $S \subseteq V$ we will write $S^{\complement}$ to denote $V \setminus S^\perp$ and $B_{S^{\complement}}$ to denote $I + A(G \setminus S^\perp)$. Observe that $B_{S^{\complement}}$ is precisely the submatrix of B obtained by deleting the rows and columns indexed by $S^\perp$. For a given vector of variables $x = (x_1, \ldots, x_n)$ we will write Σ_S as shorthand for $\sum_{i \in S} x_i$ and x_S to denote the vector of variables indexed by the elements of S. Finally, given two homogeneous polynomials $p(x), q(x)$ of degree $2 + r$, we write $p(x) \sqsupseteq q(x)$ to indicate that $p(x) - q(x) \in \mathcal{E}_n^r$.

In what follows we use the recursive characterization of $\mathcal{E}_n^r$ given in Proposition 9.

PROPOSITION 9. *For all $r > 0$,*

$$\mathcal{E}_n^r = \left\{ \sum_{i=1}^n x_i p_i(x) : p_i(x) \in \mathcal{E}_n^{r-1},\ i = 1, \ldots, n \right\}.$$

Proof. The inclusion "$\supseteq$" is immediate. The inclusion "$\subseteq$" follows by induction on r using the following identity:

$$\begin{aligned}
\sum_{\beta \in \mathbb{N}^n,\, |\beta|=r} x^\beta x^{\mathrm{T}} B_\beta\, x &= \sum_{\beta \in \mathbb{N}^n,\, |\beta|=r} \frac{1}{r} \sum_{i=1}^n \beta_i x^\beta x^{\mathrm{T}} B_\beta\, x \\
&= \sum_{\beta \in \mathbb{N}^n,\, |\beta|=r} \frac{1}{r} \sum_{i=1}^n \beta_i x_i x^{\beta - e_i} x^{\mathrm{T}} B_\beta\, x \\
&= \frac{1}{r} \sum_{i=1}^n x_i \sum_{\beta \in \mathbb{N}^n,\, |\beta|=r} \beta_i x^{\beta - e_i} x^{\mathrm{T}} B_\beta\, x \\
&= \frac{1}{r} \sum_{i=1}^n x_i \sum_{\beta \in \mathbb{N}^n,\, |\beta|=r,\, \beta_i \geq 1} \beta_i x^{\beta - e_i} x^{\mathrm{T}} B_\beta\, x \\
&= \frac{1}{r} \sum_{i=1}^n x_i \sum_{\gamma \in \mathbb{N}^n,\, |\gamma|=r-1} (\gamma_i + 1) x^\gamma x^{\mathrm{T}} B_{\gamma + e_i}\, x,
\end{aligned}$$

where $e_i \in \mathbb{N}^n$ denotes the vector with 1 in the ith coordinate and 0 in all the other coordinates. ∎

LEMMA 10. *For $\lambda > 1$*

$$\sum_{v \in V} x_v x^{\mathrm{T}} (\lambda B - e e^{\mathrm{T}}) x \sqsupseteq \sum_{v \in V} x_v P_v(x),$$

where

$$P_v(x) = \frac{1}{\lambda - 1} \left((\lambda - 1)\Sigma_{v^\perp} - \Sigma_{v^{\complement}} \right)^2 + \frac{\lambda}{\lambda - 1} x_{v^{\complement}}^{\mathrm{T}} \left((\lambda - 1) B_{v^{\complement}} - e e^{\mathrm{T}} \right) x_{v^{\complement}}.$$

Lemma 11 below can be seen as an extension of Lemma 10. First we introduce one more piece of notation. Assume $\lambda > k$ and $v_1, \dots, v_k \in V$ are such that $v_{j+1} \in \{v_1, \dots, v_j\}^{\underline{c}}$ for $j = 1, \dots, k-1$. Put $\vec{v} := \{v_1, \dots, v_k\}$, and define

$$p_{\vec{v}}(x) = \frac{\lambda^{k-2}}{2(\lambda-1)(\lambda-2)\cdots(\lambda-k)}\Big((\lambda-k+1)\left((\lambda-k)\Sigma_{\vec{v}^\perp} - k\Sigma_{\vec{v}^{\underline{c}}}\right)^2 + 2\lambda^2 x_{\vec{v}^{\underline{c}}}^{\mathrm{T}}\left((\lambda-k)B_{\vec{v}^{\underline{c}}} - ee^{\mathrm{T}}\right)x_{\vec{v}^{\underline{c}}}\Big).$$

In the statement of Lemma 11 below we also use the following convenient notation: Given $\vec{v} = \{v_1, \dots, v_k\}$ as above and $w \in \vec{v}^{\underline{c}}$ we write $\vec{v}, w$ to denote the set $\{v_1, \dots, v_k, w\}$.

LEMMA 11. *Assume $\lambda > k+1$ and $v_1, \dots, v_k \in V$ are such that $v_{j+1} \in \{v_1, \dots, v_j\}^{\underline{c}}$ for $j = 1, \dots, k-1$ and $\{v_1, \dots, v_k\}^{\underline{c}} \neq \emptyset$. If $k = 1, 2$ or if $k = 3, 4$ and $\lambda \leq 6$, then*

$$\Sigma_V p_{\vec{v}}(x) \sqsupseteq \sum_{w \in \vec{v}^{\underline{c}}} x_w p_{\vec{v},w}(x).$$

Proof of Theorem 6. Put

$$t = \max_{S \subseteq V \text{ stable}, |S| = r} \nu^{(0)}(G \setminus S^\perp).$$

Then for any stable set $\vec{v} := \{v_1, \dots, v_r\}$ of size r we have $\nu^{(0)}(G \setminus \vec{v}^\perp) \leq t$, i.e., $x_{\vec{v}^{\underline{c}}}^{\mathrm{T}}(tB_{\vec{v}^{\underline{c}}} - ee^{\mathrm{T}})x_{\vec{v}^{\underline{c}}} \in \mathcal{E}_n^0$. Therefore, for $\lambda := t + r$ we have $p_{\vec{v}}(x) \in \mathcal{E}_n^0$. In addition, observe that for each $v \in V$ the polynomial $P_v(x)$ defined in Lemma 10 satisfies

$$P_v(x) = p_v(x) + \frac{1}{2(\lambda-1)}\left((\lambda-1)\Sigma_{v^\perp} - \Sigma_{v^{\underline{c}}}\right)^2 \sqsupseteq p_v(x).$$

Hence from Lemmas 10 and 11 it follows that

$$\left(\sum_{v \in V} x_v\right)^r x^{\mathrm{T}}(\lambda B - ee^{\mathrm{T}})x \sqsupseteq \sum_{v_1 \in V} x_{v_1} \sum_{v_2 \in v_1^{\underline{c}}} x_{v_2} \cdots \sum_{v_r \in \{v_1, \dots, v_{r-1}\}^{\underline{c}}} x_{v_r} p_{\vec{v}}(x) \in \mathcal{E}_n^r$$

and so

$$\nu^{(r)}(G) \leq \lambda = r + t = r + \max_{S \subseteq V \text{ stable}, |S| = r} \nu^{(0)}(G \setminus S^\perp). \qquad \square$$

Proof of Theorem 8. If $u \in V$ is such that $u^\perp$ induces a clique in G, then

$$\begin{aligned} x^{\mathrm{T}}(\lambda B - ee^{\mathrm{T}})x &\sqsupseteq \lambda\left(\Sigma_{u^\perp}^2 + x_{u^{\underline{c}}}^{\mathrm{T}} B_{u^{\underline{c}}} x_{u^{\underline{c}}}\right) - \Sigma_V^2 \\ &= \lambda\left(\Sigma_{u^\perp}^2 + x_{u^{\underline{c}}}^{\mathrm{T}} B_{u^{\underline{c}}} x_{u^{\underline{c}}}\right) - \left(\Sigma_{u^\perp} + \Sigma_{u^{\underline{c}}}\right)^2 \\ &= (\lambda-1)\Sigma_{u^\perp}^2 - 2\Sigma_{u^\perp}\Sigma_{u^{\underline{c}}} + \tfrac{1}{\lambda-1}\Sigma_{u^{\underline{c}}}^2 + \lambda x_{u^{\underline{c}}}^{\mathrm{T}} B_{u^{\underline{c}}} x_{u^{\underline{c}}} - \tfrac{\lambda}{\lambda-1}\Sigma_{u^{\underline{c}}}^2 \\ &= P_u(x) \\ &\sqsupseteq p_u(x). \end{aligned}$$

Hence by taking $v_1 := u$, we can modify the last step in the proof of Theorem 6 to

$$\left(\sum_{v \in V} x_v\right)^{r-1} x^{\mathrm{T}}(\lambda B - ee^{\mathrm{T}})x \sqsupseteq \sum_{v_2 \in v_1^{\underline{c}}} x_{v_2} \cdots \sum_{v_r \in \{v_1, \dots, v_{r-1}\}^{\underline{c}}} x_{v_r} p_{\vec{v}}(x) \in \mathcal{E}_n^{r-1},$$

as long as $\lambda := r + t$, where

$$t := \max_{S \subseteq V \text{ stable } |S|=r,\, u \in S} \nu^{(0)}(G \setminus S^{\perp}).$$

Hence we obtain

$$\nu^{(r-1)}(G) \leq \lambda = r + t = r + \max_{S \subseteq V \text{ stable } |S|=r,\, u \in S} \nu^{(0)}(G \setminus S^{\perp}). \qquad \square$$

We conclude this section with the proofs of Lemmas 10 and 11. The proof of Lemma 10 relies on the following lemma.

LEMMA 12.

$$\sum_{v \in V} x_v \Sigma_{v^{\perp}} \Sigma_{v^{\subseteq}} \sqsupseteq \sum_{v \in V} x_v x_{v^{\subseteq}}^{\mathrm{T}} B_{v^{\subseteq}} x_{v^{\subseteq}}.$$

Proof.

$$\begin{aligned}
\sum_{v \in V} x_v \Sigma_{v^{\perp}} \Sigma_{v^{\subseteq}} &= \sum_{v \in V} \sum_{u \in v^{\subseteq}} x_u x_v \Sigma_{v^{\perp}} \\
&= \sum_{u \in V} \sum_{v \in u^{\subseteq}} x_u x_v \Sigma_{v^{\perp}} \\
&\sqsupseteq \sum_{u \in V} x_u \sum_{v \in u^{\subseteq}} x_v \sum_{w \in v^{\perp} \cap u^{\subseteq}} x_w \\
&= \sum_{u \in V} x_u x_{u^{\subseteq}}^{\mathrm{T}} B_{u^{\subseteq}} x_{u^{\subseteq}}. \qquad \square
\end{aligned}$$

The proofs below also rely on the following observation.

Observation 13. $x^{\mathrm{T}}(\lambda B - ee^{\mathrm{T}})x = \sum_{v \in V} x_v((\lambda - 1)\Sigma_{v^{\perp}} - \Sigma_{v^{\subseteq}})$.

Proof of Lemma 10. From Observation 13 and Lemma 12 it follows that

$$\begin{aligned}
\sum_{v \in V} x_v x^{\mathrm{T}} \left(\lambda B - ee^{\mathrm{T}}\right) x &= \sum_{v \in V} x_v((\lambda - 1)\Sigma_{v^{\perp}} - \Sigma_{v^{\subseteq}})\Sigma_V \\
&= \sum_{v \in V} x_v \left(\frac{1}{\lambda - 1} \left((\lambda - 1)\Sigma_{v^{\perp}} - \Sigma_{v^{\subseteq}}\right)^2 + \frac{\lambda}{\lambda - 1} \left((\lambda - 1)\Sigma_{v^{\perp}} - \Sigma_{v^{\subseteq}}\right) \Sigma_{v^{\subseteq}} \right) \\
&\sqsupseteq \sum_{v \in V} x_v \left(\frac{1}{\lambda - 1} \left((\lambda - 1)\Sigma_{v^{\perp}} - \Sigma_{v^{\subseteq}}\right)^2 + \frac{\lambda}{\lambda - 1} x_{v^{\subseteq}}^{\mathrm{T}} \left((\lambda - 1)B_{v^{\subseteq}} - ee^{\mathrm{T}}\right) x_{v^{\subseteq}} \right) \\
&= \sum_{v \in V} x_v P_v(x). \qquad \square
\end{aligned}$$

Proof of Lemma 11. Put

$$C := \frac{\lambda^{k-2}}{2(\lambda - 1)(\lambda - 2) \cdots (\lambda - k)}.$$

By dropping some terms and the fact that $\Sigma_V = \Sigma_{\vec{v}^{\subseteq}} + \Sigma_{\vec{v}^{\perp}}$, it follows that

$$\begin{aligned}
\Sigma_V p_{\vec{v}}(x) \sqsupseteq C \Big(& (\lambda - k + 1)\Sigma_{\vec{v}^{\subseteq}} \left((\lambda - k)\Sigma_{\vec{v}^{\perp}} - k\Sigma_{\vec{v}^{\subseteq}}\right)^2 \\
& + 2\lambda^2 \Sigma_{\vec{v}^{\subseteq}} x_{\vec{v}^{\subseteq}}^{\mathrm{T}} \left((\lambda - k)B_{\vec{v}^{\subseteq}} - ee^{\mathrm{T}}\right) x_{\vec{v}^{\subseteq}} \\
& + 2\lambda^2 \Sigma_{\vec{v}^{\perp}} x_{\vec{v}^{\subseteq}}^{\mathrm{T}} \left((\lambda - k)B_{\vec{v}^{\subseteq}} - ee^{\mathrm{T}}\right) x_{\vec{v}^{\subseteq}} \Big).
\end{aligned}$$

To simplify notation, put $\ell := \lambda - k$. Then the latter inequality can be rewritten as

$$\begin{aligned}\Sigma_V p_{\vec{v}}(x) \sqsupseteq C\Big(& (\ell+1)\Sigma_{\vec{v}^{\subseteq}} \left(\ell\, \Sigma_{\vec{v}^{\perp}} - k\Sigma_{\vec{v}^{\subseteq}}\right)^2 \\ & + \; 2\lambda^2 \Sigma_{\vec{v}^{\subseteq}} x_{\vec{v}^{\subseteq}}^{\mathrm{T}} \left(\ell\, B_{\vec{v}^{\subseteq}} - ee^{\mathrm{T}}\right) x_{\vec{v}^{\subseteq}} \\ & + \; 2\lambda^2 \Sigma_{\vec{v}^{\perp}} x_{\vec{v}^{\subseteq}}^{\mathrm{T}} \left(\ell\, B_{\vec{v}^{\subseteq}} - ee^{\mathrm{T}}\right) x_{\vec{v}^{\subseteq}} \Big).\end{aligned} \tag{18}$$

By Observation 13 (applied to the graph $G := G \setminus \vec{v}^{\perp}$ and $\lambda := \ell$) and the fact that $\Sigma_{\vec{v}^{\perp}} = \frac{1}{\ell}\left(\ell\, \Sigma_{\vec{v}^{\perp}} - k\Sigma_{\vec{v}^{\subseteq}}\right) + \frac{k}{\ell}\Sigma_{\vec{v}^{\subseteq}}$, we can rewrite $\Sigma_{\vec{v}^{\perp}} x_{\vec{v}^{\subseteq}}^{\mathrm{T}} \left(\ell\, B_{\vec{v}^{\subseteq}} - ee^{\mathrm{T}}\right) x_{\vec{v}^{\subseteq}}$ as

$$\frac{1}{\ell} \sum_{w \in \vec{v}^{\subseteq}} x_w \left(\ell\, \Sigma_{\vec{v}^{\perp}} - k\Sigma_{\vec{v}^{\subseteq}}\right) \left((\ell-1)\Sigma_{w^{\perp} \cap \vec{v}^{\subseteq}} - \Sigma_{\{\vec{v},w\}^{\subseteq}}\right) + \frac{k}{\ell}\Sigma_{\vec{v}^{\subseteq}} x_{\vec{v}^{\subseteq}}^{\mathrm{T}} \left(\ell\, B_{\vec{v}^{\subseteq}} - ee^{\mathrm{T}}\right) x_{\vec{v}^{\subseteq}}.$$

Thus from (18) we get

$$\Sigma_V p_{\vec{v}}(x) \sqsupseteq C \sum_{w \in \vec{v}^{\subseteq}} x_w Q_{\vec{v},w}(x), \tag{19}$$

where

$$\begin{aligned} Q_{\vec{v},w}(x) = & (\ell+1)\left(\ell\, \Sigma_{\vec{v}^{\perp}} - k\Sigma_{\vec{v}^{\subseteq}}\right)^2 \\ & + \; \tfrac{2\lambda^3}{\ell} x_{\vec{v}^{\subseteq}}^{\mathrm{T}} \left(\ell\, B_{\vec{v}^{\subseteq}} - ee^{\mathrm{T}}\right) x_{\vec{v}^{\subseteq}} \\ & + \; \tfrac{2\lambda^2}{\ell} \left(\ell\, \Sigma_{\vec{v}^{\perp}} - k\Sigma_{\vec{v}^{\subseteq}}\right) \left((\ell-1)\Sigma_{w^{\perp} \cap \vec{v}^{\subseteq}} - \Sigma_{\{\vec{v},w\}^{\subseteq}}\right). \end{aligned}$$

By Lemma 10 (applied to the graph $G := G \setminus \vec{v}^{\perp}$ and $\lambda := \ell$), it follows that

$$\sum_{w \in \vec{v}^{\subseteq}} x_w Q_{\vec{v},w}(x) \sqsupseteq \sum_{w \in \vec{v}^{\subseteq}} x_w q_{\vec{v},w}(x), \tag{20}$$

where

$$\begin{aligned} q_{\vec{v},w}(x) = & (\ell+1)\left(\ell\, \Sigma_{\vec{v}^{\perp}} - k\Sigma_{\vec{v}^{\subseteq}}\right)^2 \\ & + \; \tfrac{2\lambda^3}{\ell(\ell-1)} \left((\ell-1)\, \Sigma_{w^{\perp} \cap \vec{v}^{\subseteq}} - \Sigma_{\{\vec{v},w\}^{\subseteq}}\right)^2 \\ & + \; \tfrac{2\lambda^3}{\ell-1} x_{\{\vec{v},w\}^{\subseteq}}^{\mathrm{T}} \left((\ell-1)B_{\{\vec{v},w\}^{\subseteq}} - ee^{\mathrm{T}}\right) x_{\{\vec{v},w\}^{\subseteq}} \\ & + \; \tfrac{2\lambda^2}{\ell} \left(\ell\, \Sigma_{\vec{v}^{\perp}} - k\Sigma_{\vec{v}^{\subseteq}}\right) \left((\ell-1)\Sigma_{w^{\perp} \cap \vec{v}^{\subseteq}} - \Sigma_{\{\vec{v},w\}^{\subseteq}}\right). \end{aligned}$$

Using some algebraic manipulations and the fact that $\ell = \lambda - k$, we can rewrite $q_{\vec{v},w}(x)$ as

$$\begin{aligned} q_{\vec{v},w}(x) = & \tfrac{\lambda(\ell-1)}{\ell} \left(\ell\, \Sigma_{\vec{v}^{\perp}} - k\Sigma_{\vec{v}^{\subseteq}} + \tfrac{\lambda}{\ell-1} \left((\ell-1)\Sigma_{w^{\perp} \cap \vec{v}^{\subseteq}} - \Sigma_{\{\vec{v},w\}^{\subseteq}}\right)\right)^2 \\ & + \; \tfrac{\lambda(2-k)+k^2-k}{\ell} \left(\ell\, \Sigma_{\vec{v}^{\perp}} - k\Sigma_{\vec{v}^{\subseteq}}\right)^2 \\ & + \; \tfrac{\lambda^3}{\ell(\ell-1)} \left((\ell-1)\Sigma_{w^{\perp} \cap \vec{v}^{\subseteq}} - \Sigma_{\{\vec{v},w\}^{\subseteq}}\right)^2 \\ & + \; \tfrac{2\lambda^3}{\ell-1} x_{\{\vec{v},w\}^{\subseteq}}^{\mathrm{T}} \left((\ell-1)B_{\{\vec{v},w\}^{\subseteq}} - ee^{\mathrm{T}}\right) x_{\{\vec{v},w\}^{\subseteq}}. \end{aligned} \tag{21}$$

Since $\Sigma_{\vec{v}^{\subseteq}} = \Sigma_{(w^{\perp} \cap \vec{v}^{\subseteq}) \cup \{\vec{v},w\}^{\subseteq}} = \Sigma_{w^{\perp} \cap \vec{v}^{\subseteq}} + \Sigma_{\{\vec{v},w\}^{\subseteq}}$, it follows that

$$\begin{aligned} & \left(\ell\, \Sigma_{\vec{v}^{\perp}} - k\Sigma_{\vec{v}^{\subseteq}} + \tfrac{\lambda}{\ell-1} \left((\ell-1)\Sigma_{w^{\perp} \cap \vec{v}^{\subseteq}} - \Sigma_{\{\vec{v},w\}^{\subseteq}}\right)\right)^2 = \\ & \qquad \left(\ell\, \Sigma_{\vec{v}^{\perp}} + (\lambda-k)\Sigma_{w^{\perp} \cap \vec{v}^{\subseteq}} - \tfrac{k\ell+\lambda-k}{\ell-1}\Sigma_{\{\vec{v},w\}^{\subseteq}}\right)^2 = \\ & \qquad\qquad \left(\ell\, \Sigma_{\{\vec{v},w\}^{\perp}} - \tfrac{(k+1)\ell}{\ell-1}\Sigma_{\{\vec{v},w\}^{\subseteq}}\right)^2. \end{aligned} \tag{22}$$

The last step holds because $\ell = \lambda - k$ and $\Sigma_{\{\vec{v},w\}^{\perp}} = \Sigma_{\vec{v}^{\perp}} + \Sigma_{w^{\perp} \cap \vec{v}^{\subseteq}}$.

Since $\lambda > k+1$, it follows that $\lambda(2-k)+k^2-k \geq 0$ if $k = 1, 2$, or if $k = 3, 4$ and $\lambda \leq 6$. Thus by dropping the second and third terms in (21) and using (22), we get

$$
\begin{aligned}
q_{\vec{v},w}(x) &\sqsupseteq \tfrac{\lambda(\ell-1)}{\ell}\left(\ell\,\Sigma_{\{\vec{v},w\}^\perp} - \tfrac{(k+1)\ell}{\ell-1}\Sigma_{\{\vec{v},w\}^{\subseteq}}\right)^2 \\
&\quad + \tfrac{2\lambda^3}{\ell-1}x^{\mathrm{T}}_{\{\vec{v},w\}^{\subseteq}}\left((\ell-1)B_{\{\vec{v},w\}^{\subseteq}} - ee^{\mathrm{T}}\right)x_{\{\vec{v},w\}^{\subseteq}} \\
&= \tfrac{\lambda\ell}{\ell-1}\left((\ell-1)\Sigma_{\{\vec{v},w\}^\perp} - (k+1)\Sigma_{\{\vec{v},w\}^{\subseteq}}\right)^2 \\
&\quad + \tfrac{2\lambda^3}{\ell-1}x^{\mathrm{T}}_{\{\vec{v},w\}^{\subseteq}}\left((\ell-1)B_{\{\vec{v},w\}^{\subseteq}} - ee^{\mathrm{T}}\right)x_{\{\vec{v},w\}^{\subseteq}} \\
&= \tfrac{\lambda(\lambda-k)}{\lambda-k-1}\left((\lambda-k-1)\Sigma_{\{\vec{v},w\}^\perp} - (k+1)\Sigma_{\{\vec{v},w\}^{\subseteq}}\right)^2 \qquad (23)\\
&\quad + \tfrac{2\lambda^3}{\lambda-k-1}x^{\mathrm{T}}_{\{\vec{v},w\}^{\subseteq}}\left((\lambda-k-1)B_{\{\vec{v},w\}^{\subseteq}} - ee^{\mathrm{T}}\right)x_{\{\vec{v},w\}^{\subseteq}} \\
&= \tfrac{\lambda}{\lambda-k-1}\Big((\lambda-k)\left((\lambda-k-1)\Sigma_{\{\vec{v},w\}^\perp} - (k+1)\Sigma_{\{\vec{v},w\}^{\subseteq}}\right)^2 \\
&\qquad\qquad + 2\lambda^2 x^{\mathrm{T}}_{\{\vec{v},w\}^{\subseteq}}\left((\lambda-k-1)B_{\{\vec{v},w\}^{\subseteq}} - ee^{\mathrm{T}}\right)x_{\{\vec{v},w\}^{\subseteq}}\Big) \\
&= \tfrac{p_{\vec{v},w}(x)}{C}.
\end{aligned}
$$

Combining (19), (20), and (23) we get

$$\Sigma_V p_{\vec{v}}(x) \sqsupseteq \sum_{w\in\vec{v}^{\subseteq}} x_w p_{\vec{v},w}(x). \qquad \square$$

5. Some special classes of graphs. Recall that the *chromatic number* $\chi(G)$ of a graph $G = (V, E)$ is the minimum number of colors required to color the vertices of G so that for all $\{i, j\} \in E$ the vertices i and j have different colors. Recall also that the *complement* $\overline{G}$ of $G = (V, E)$ is the graph with the same vertex set V and set of edges $\{\{i, j\} : \{i, j\} \notin E\}$.

We next discuss in further detail three important classes of graphs: the class of graphs G with $\alpha(G) = \chi(\overline{G})$, the cycles C_n, and their complements $\overline{C}_n$. For the first class of graphs, Corollary 15 below shows that $\nu^{(0)}(\cdot) = \alpha(\cdot)$. This readily applies to even cycles and their complements. By contrast, $\nu^{(0)}(\cdot) > \alpha(\cdot)$ for odd cycles and their complements. Furthermore, for these classes of graphs, the identities in Examples 19 and 21 show that $\nu^{(1)}(\cdot) = \alpha(\cdot)$.

5.1. Graphs G with $\chi(\overline{G}) = \alpha(G)$. Proposition 14 below gives a succinct proof of the inequality $\nu^{(0)}(G) \leq \chi(\overline{G})$. We note that this inequality also follows from putting together the Lovász sandwich theorem [11], the identity $\nu^{(0)}(G) = \vartheta^{(0)}(G) = \vartheta'(G)$ due to de Klerk and Pasechnik [4], and the inequality $\vartheta'(G) \leq \vartheta(G)$ due to Schrijver [15]. Here $\vartheta'(\cdot)$ and $\vartheta(\cdot)$ are, respectively, Schrijver's ϑ' function [15] and Lovász ϑ function [11]. However, the proof below is direct and provides an interesting constructive procedure. This procedure will allow us to derive the identities in Examples 17, 18, and 19.

PROPOSITION 14. $\nu^{(0)}(G) \leq \chi(\overline{G})$.

Proof. To simplify notation, we shall write χ as shorthand for $\chi(\overline{G})$. Fix a coloring of $\overline{G}$ with χ colors, and let V_j be the set of vertices colored with color j for $j = 1, \ldots, \chi$. Since each V_j is a clique in G,

$$x^{\mathrm{T}}(I + A(G))x = \sum_{j=1}^{\chi}\left(\sum_{i\in V_j} x_i\right)^2 + q(x) \qquad (24)$$

for some quadratic form $q(x)$ with nonnegative coefficients.

On the other hand,

$$\chi \cdot \sum_{j=1}^{\chi} \left(\sum_{i \in V_j} x_i \right)^2 - \left(\sum_{i \in V} x_i \right)^2 = \sum_{1 \le j < k \le \chi} \left(\sum_{i \in V_j} x_i - \sum_{i \in V_k} x_i \right)^2. \tag{25}$$

From (24) and (25) we get

$$x^{\mathrm{T}} \left(\chi \cdot (I + A(G)) - ee^{\mathrm{T}} \right) x = \sum_{1 \le j < k \le \chi} \left(\sum_{i \in V_j} x_i - \sum_{i \in V_k} x_i \right)^2 + \chi \cdot q(x).$$

Thus $\chi \cdot (I + A(G)) - ee^{\mathrm{T}} \in \mathcal{Q}_n^0$, and consequently $\nu^{(0)}(G) \le \chi(\overline{G})$. □

COROLLARY 15. *If $\alpha(G) = \chi(\overline{G})$, then $\nu^{(0)}(G) = \alpha(G)$.*

From Corollary 15, it readily follows that $\nu^{(0)}(C_n) = \alpha(C_n)$ and $\nu^{(0)}(\overline{C}_n) = \alpha(\overline{C}_n)$ for n even or $n = 3$. By contrast, we next show that $\nu^{(0)}(C_n) > \alpha(C_n) = \nu^{(1)}(C_n)$ and $\nu^{(0)}(\overline{C}_n) > \alpha(\overline{C}_n) = \nu^{(1)}(\overline{C}_n)$ for $n \ge 5$ and odd.

5.2. Odd cycles. For ease of notation, throughout this section the arithmetic operations in the indices below are meant to be performed modulo $2m + 1$, i.e., $1 + 2m + 1 = 1$, $3 + 2m = 2$, etc.

PROPOSITION 16. *Assume $m \ge 2$. Then $\nu^{(0)}(C_{2m+1}) > m = \alpha(C_{2m+1})$.*

Proof. Assume $\nu^{(0)}(C_{2m+1}) = m$. Then $m(I + A(C_{2m+1})) - ee^{\mathrm{T}} \in \mathcal{Q}_n^0$, i.e.,

$$m(I + A(C_{2m+1})) - ee^{\mathrm{T}} = RR + N$$

for some $R, N \in \mathbb{S}^n$ with $N \ge 0$. Let R_j denote the jth column of R. It follows that for any stable set $S \subseteq V$ of size m

$$0 = \left\| \sum_{j \in S} R_j \right\|^2 + \sum_{i,j \in S} N_{ij} \ge \left\| \sum_{j \in S} R_j \right\|^2,$$

and thus $\sum_{j \in S} R_j = 0$. Taking the stable sets $S = \{j, j+3, j+5, \ldots, j+2m-1\}$ and $S' = \{j+1, j+3, j+5, \ldots, j+2m-1\}$, we get $0 = R_j + \sum_{k=1}^{m-1} R_{j+2k+1} = R_{j+1} + \sum_{k=1}^{m-1} R_{j+2k+1}$ and so $R_j = R_{j+1}$ for all j. Applying this to $j = 1, 2, \ldots, 2m+1$, we conclude that $R = 0$. But this yields $m(I + A(C_{2m+1})) - ee^{\mathrm{T}} = N \ge 0$, which is clearly a contradiction. □

Examples 17, 18, and 19 below show that $m(I + A(C_{2m+1})) - ee^{\mathrm{T}} \in \mathcal{Q}_n^1$ for every odd cycle C_{2m+1} with $m \ge 2$. This in particular yields $\nu^{(1)}(C_{2m+1}) = m = \alpha(C_{2m+1})$. The identities in Examples 17, 18, and 19 follow from specializing the proofs of Theorem 6 and Proposition 14 to odd and even cycles, respectively. We note that the identity in Example 17 had been derived by Parrilo [14, Chapter 5].

Example 17. Let $p(x_1, x_2, x_3, x_4, x_5) = (x_1 + x_2 + x_5 - x_3 - x_4)^2 + 4x_2x_3$. Then

$$\sum_{i=1}^{5} x_i x^{\mathrm{T}} \left(2(I + A(C_5)) - ee^{\mathrm{T}} \right) x = \sum_{i=1}^{5} x_i p(x_i, x_{i+1}, x_{i+2}, x_{i+3}, x_{i+4}).$$

Example 18. Let

$$p(x_1, \ldots, x_7) = 2 \left(x_1 + x_2 + x_7 - \frac{1}{2}(x_3 + x_4 + x_5 + x_6) \right)^2 + \frac{3}{2}(x_3 + x_4 - x_5 - x_6)^2 + 6(x_2x_3 + x_4x_5).$$

Then

$$\sum_{i=1}^{7} x_i x^{\mathrm{T}} \left(3(I + A(C_7)) - ee^{\mathrm{T}}\right) x = \sum_{i=1}^{7} x_i p(x_i, x_{i+1}, \ldots, x_{i+6}).$$

Example 19. Assume $m \geq 2$. Let

$$\begin{aligned} p(x) = (m-1) &\left(x_1 + x_2 + x_{2m+1} - \frac{1}{m-1} \sum_{k=3}^{2m} x_k \right)^2 \\ &+ \frac{m}{m-1} \sum_{1 \leq i < j \leq m-1} \left(x_{2i+1} + x_{2i+2} - x_{2j+1} - x_{2j+2}\right)^2 \\ &+ 2m \sum_{i=1}^{m-1} x_{2i} x_{2i+1}. \end{aligned}$$

Then

$$\sum_{i=1}^{2m+1} x_i x^{\mathrm{T}} \left(m(I + A(C_{2m+1})) - ee^{\mathrm{T}}\right) x = \sum_{i=1}^{2m+1} x_i p(x_i, x_{i+1}, \ldots, x_{i+2m}).$$

5.3. Complements of odd cycles. The following proposition is proven with an argument similar (but simpler) to that used in the proof of Proposition 16.

PROPOSITION 20. *Assume $m \geq 2$. Then $\nu^{(0)}(\overline{C}_{2m+1}) > 2 = \alpha(\overline{C}_{2m+1})$.*

Example 21 below shows that $2(I + A(\overline{C}_{2m+1})) - ee^{\mathrm{T}} \in \mathcal{Q}_n^1$ for every $\overline{C}_{2m+1}$, which in turn yields $\nu^{(1)}(\overline{C}_{2m+1}) = 2 = \alpha(\overline{C}_{2m+1})$. The identity in Example 21 follows from specializing the proof of Theorem 6 to complements of odd cycles.

Example 21. Assume $m \geq 2$. Let

$$p(x) = \left(x_1 - x_2 - x_{2m+1} + \sum_{k=3}^{2m} x_k \right)^2 + 4x_2 \sum_{k=4}^{2m} x_k.$$

Then

$$\sum_{i=1}^{2m+1} x_i x^{\mathrm{T}} \left(2(I + A(\overline{C}_{2m+1})) - ee^{\mathrm{T}}\right) x = \sum_{i=1}^{2m+1} x_i p(x_i, x_{i+1}, \ldots, x_{i+2m}).$$

5.4. The smallest graphs G such that $\nu^{(\alpha(G)-2)}(G) > \alpha(G)$ for $\alpha(G) = 2, 3, 4, 5$. The examples below show that the result in Corollary 7 concerning the attainment of $\alpha(G)$ is tight on r. Furthermore, the graphs in these examples are the smallest possible. Several of these examples rely on numerical computations for $\nu^{(r)}(G)$. These numerical results were obtained by solving the semidefinite programming formulations using `SeDuMi` [16], with a precision of at least `1E-8`.

From Theorem 6 and Proposition 16 it follows that C_5 is the smallest (with fewest vertices) graph G such that $\nu^{(0)}(G) > \alpha(G)$. In addition, again by Theorem 6

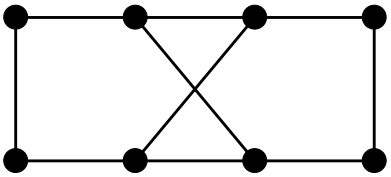
$$\nu^{(1)}(G) \leq 1 + \max_{v \in V} \nu^{(0)}(G \setminus v^{\perp}). \tag{26}$$

Furthermore, if some vertex $u \in V$ is such that $\deg(u) = 0$ or 1, then $u^{\perp}$ induces a clique in G, and Theorem 8 yields

$$\nu^{(1)}(G) \leq 2 + \max_{v \in u^{\subseteq}} \nu^{(0)}(G \setminus \{u, v\}^{\perp}). \tag{27}$$

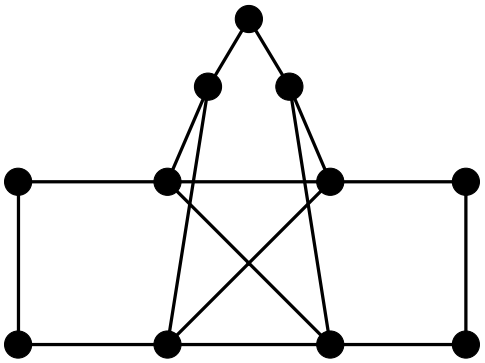
From (26) and (27) and the fact that C_5 is the smallest graph G such that $\nu^{(0)}(G) > \alpha(G)$, it follows that $\nu^{(1)}(G) = \alpha(G)$ if G has at most seven vertices and $\alpha(G) = 3$.

Now consider the following graph G_8:



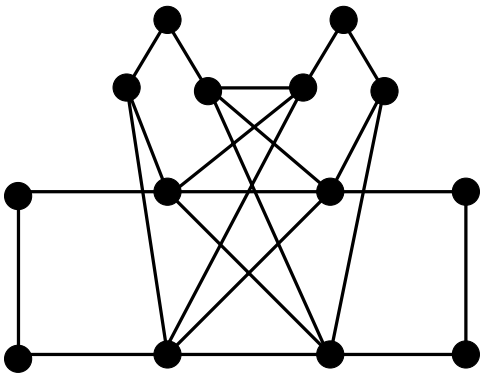
We have $\nu^{(1)}(G_8) = 3.043276 > 3 = \alpha(G_8)$. Furthermore, it is easy to show that if G' is another graph with eight vertices such that $\nu^{(1)}(G') > \alpha(G')$, then the vertices of G' can be numbered so that $E(G_8) \subseteq E(G')$. Thus G_8 is the smallest graph G such that $\nu^{(1)}(G) > \alpha(G) = 3$.

We can extend the above reasoning further. Again, from observations above, it follows that $\nu^{(2)}(G) = \alpha(G)$ if G has at most ten vertices and $\alpha(G) = 4$. Consider the following graph G_{11}:



We have $\nu^{(2)}(G_{11}) = 4.011111 > 4 = \alpha(G_{11})$.

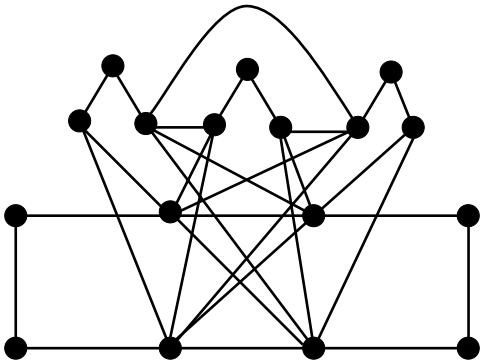
We can go yet one step further. Again, from observations above, it follows that $\nu^{(3)}(G) = \alpha(G)$ if G has at most 13 vertices and $\alpha(G) = 5$. Now Consider the following graph G_{14}:



We have $\nu^{(3)}(G_{14}) = 5.004886 > 5 = \alpha(G_{14})$.

It is interesting to note that (up to the a numerical accuracy of `1E-8`) both $\vartheta^{(2)}(G_{11}) = \alpha(G_{11})$ and $\vartheta^{(2)}(G_{14}) = \alpha(G_{14})$. So in general $\nu^{(r)}(\cdot) > \vartheta^{(r)}(\cdot)$ for $r > 1$ even though $\nu^{(r)}(\cdot) = \vartheta^{(r)}(\cdot)$ for $r = 0, 1$.

We conjecture that the general construction suggested above gives graphs G with $\alpha(G) = r$ and $\nu^{(r-2)}(G) > r$ for any r. For example, for $r = 4$ consider the following graph G_{17}:



It should be the case that $\nu^{(4)}(G_{17}) > \alpha(G_{17})$. Unfortunately the semidefinite program involved in the calculation of $\nu^{(4)}(G_{17})$ is beyond our current computational capabilities. For this graph we did find that $\vartheta^{(2)}(G_{17}) = 6.000475 > 6 = \alpha(G_{17})$. To this date this is the smallest explicit example of a graph G with $\vartheta^{(2)}(G) > \alpha(G)$.

Acknowledgment. We thank the anonymous referees whose helpful suggestions substantially improved the paper.

REFERENCES

[1] E. Balas, S. Ceria, and G. Cornuéjols, *A lift-and-project cutting plane algorithm for mixed* 0-1 *programs*, Math. Program., 58 (1993), pp. 295–324.

[2] I. Bomze, M. Budinich, P. Pardalos, and M. Pelillo, *The Maximum Clique Problem*, in Handb. Combin. Opt., Suppl. A, D. Du and P. Pardalos, eds., Kluwer Academic Publishers, Norwell, MA, 1999.

[3] I. Bomze and E. de Klerk, *Solving standard quadratic optimization problems via semidefinite and copositive programming*, J. Global Optim., 24 (2002), pp. 163–185.

[4] E. de Klerk and D. Pasechnik, *Approximation of the stability number of a graph via copositive programming*, SIAM J. Optim., 12 (2002), pp. 875–892.

[5] P. Diananda, *On non-negative forms in real variables some or all of which are non-negative*, Math. Proc. Cambridge Philos. Soc., 58 (1962), pp. 17–25.

[6] N. Gvozdenović and M. Laurent, *Semidefinite bounds for the stability number of a graph via sums of squares of polynomials*, in Integer Programming and Combinatorial Optimization, Lecture Notes in Comput. Sci. 3509/2005, Springer, Berlin, 2005, pp. 136–151.

[7] J. Håstad, *Clique is hard to approximate within* $|V^{1-\epsilon}|$, Acta Math., 182 (1999), pp. 105–142.

[8] D. Hochbaum (ed), *Approximation Algorithms for NP-Hard Problems*, PWS Publishing Company, Boston, MA, 1998.

[9] J. Lasserre, *Global optimization problems with polynomials and the problem of moments*, SIAM J. Optim., 11 (2001), pp. 796–817.

[10] J. Lasserre, *An explicit equivalent positive semidefinite program for nonlinear* 0-1 *programs*, SIAM J. Optim., 12 (2002), pp. 756–769.

[11] L. Lovász, *On the Shannon capacity of a graph*, IEEE Trans. Inform. Theory, 25 (1979), pp. 1–7.

[12] L. Lovász and A. Schrijver, *Cones of matrices and set-functions and* 0-1 *optimization*, SIAM J. Optim., 1 (1991), pp. 166–190.

[13] K. Murty and S. Kabadi, *Some NP-complete problems in quadratic and linear programming*, Math. Program., 39 (1987), pp. 117–129.
[14] P. Parrilo, *Structured Semidefinite Programming and Algebraic Geometry Methods in Robustness and Optimization,* Ph.D. thesis, California Institute of Technology, Pasadena, CA, 2000.
[15] A. Schrijver, *A comparison of Delsarte and Lovász bounds*, IEEE Trans. Inform. Theory, 25 (1979), pp. 425–429.
[16] J. Sturm, *Using SeDuMi 1.02, a Matlab toolbox for optimization over symmetric cones*, Optim. Methods Softw., 11-12 (1999), pp. 545–581.
[17] L. Zuluaga, J. Vera, and J. Peña, *LMI approximations for cones of positive semidefinite forms*, SIAM J. Optim., 16 (2006), 1076–1091.

SIAM J. OPTIM.
Vol. 18, No. 1, pp. 106–132

APPROXIMATE GAUSS–NEWTON METHODS FOR NONLINEAR LEAST SQUARES PROBLEMS*

S. GRATTON†, A. S. LAWLESS‡, AND N. K. NICHOLS‡

Abstract. The Gauss–Newton algorithm is an iterative method regularly used for solving nonlinear least squares problems. It is particularly well suited to the treatment of very large scale variational data assimilation problems that arise in atmosphere and ocean forecasting. The procedure consists of a sequence of linear least squares approximations to the nonlinear problem, each of which is solved by an "inner" direct or iterative process. In comparison with Newton's method and its variants, the algorithm is attractive because it does not require the evaluation of second-order derivatives in the Hessian of the objective function. In practice the exact Gauss–Newton method is too expensive to apply operationally in meteorological forecasting, and various approximations are made in order to reduce computational costs and to solve the problems in real time. Here we investigate the effects on the convergence of the Gauss–Newton method of two types of approximation used commonly in data assimilation. First, we examine "truncated" Gauss–Newton methods where the inner linear least squares problem is not solved exactly, and second, we examine "perturbed" Gauss–Newton methods where the true linearized inner problem is approximated by a simplified, or perturbed, linear least squares problem. We give conditions ensuring that the truncated and perturbed Gauss–Newton methods converge and also derive rates of convergence for the iterations. The results are illustrated by a simple numerical example. A practical application to the problem of data assimilation in a typical meteorological system is presented.

Key words. nonlinear least squares problems, approximate Gauss–Newton methods, variational data assimilation

AMS subject classifications. 65K10, 65H10, 90C30, 90C06

DOI. 10.1137/050624935

1. Introduction.

The Gauss–Newton (GN) method is a well-known iterative technique used regularly for solving the nonlinear least squares problem (NLSP)

$$\min_x \phi(x) = \frac{1}{2} \|f(x)\|_2^2, \tag{1}$$

where x is an n-dimensional real vector and f is an m-dimensional real vector function of x [20].

Problems of this form arise commonly from applications in optimal control and filtering and in data fitting. As a simple example, if we are given m observed data (t_i, y_i) that we wish to fit with a model $S(x,t)$, determined by a vector x of n parameters, and if we define the ith component of $f(x)$ to be $f_i(x) = S(x,t_i) - y_i$, then the solution to the NLSP (1) gives the best model fit to the data in the sense of the minimum sum of square errors. The choice of norm is often justified by statistical considerations [22].

Recently, very large inverse problems of this type arising in *data assimilation* for numerical weather, ocean, and climate prediction and for other applications in the environmental sciences have attracted considerable attention [11, 6, 18, 19, 14].

*Received by the editors February 21, 2005; accepted for publication (in revised form) October 16, 2006; published electronically February 2, 2007.
http://www.siam.org/journals/siopt/18-1/62493.html
†CERFACS, 42 Avenue Gustave Coriolis, 31057 Toulouse, CEDEX, France (s.gratton@cerfacs.fr).
‡Department of Mathematics, The University of Reading, P.O. Box 220, Reading, RG6 6AX United Kingdom (a.s.lawless@reading.ac.uk, n.k.nichols@reading.ac.uk). The work of the third author was in part supported by the U.S. Department of Energy.

In data assimilation a set of observed data is matched to the solution of a discrete dynamical model of a physical system over a period of time. The aim is to provide a "best" estimate of the current state of the system to enable accurate forecasts to be made of the future system behavior. Operationally the incremental four-dimensional variational data assimilation technique (4D-Var) is now used in many meteorological forecasting centers [6]. Recently it has been established that this method corresponds to a GN procedure [14, 15].

The GN method consists of solving a sequence of linearized least squares approximations to the nonlinear (NLSP) problem, each of which can be solved efficiently by an "inner" direct or iterative process. In comparison with Newton's method and its variants, the GN method for solving the NLSP is attractive because it does not require computation or estimation of the second derivatives of the function $f(x)$ and hence is numerically more efficient.

In practice, particularly for the very large problems arising in data assimilation, approximations are made within the GN process in order to reduce computational costs. The effects of these approximations on the convergence of the method need to be understood. Here we investigate the effects of two types of approximation used commonly in data assimilation: First, we examine "truncated" GN (TGN) methods where the inner linear least squares problem is not solved exactly, and second, we examine "perturbed" GN (PGN) methods where the true linearized inner problem is approximated by a simplified, or perturbed, linear least squares problem. We give conditions ensuring that the truncated and perturbed GN methods converge and also derive rates of convergence for the iterations.

In the next section we state the problem in detail, together with our assumptions, and define the GN algorithm. We also present some basic theory for the exact method. The truncated and perturbed algorithms that are to be investigated are then defined. In the following sections theoretical convergence results are established for the approximate GN methods. Two different approaches are used to derive the theory. First, we apply extensions of the results of [20, 8] for inexact Newton (IN) methods to the approximate GN methods in order to obtain general convergence theorems. We then derive more restricted results using the approach of [9]. The restricted results also provide estimates for the rates of convergence of the methods. Conditions for linear, superlinear, and quadratic convergence are noted. Numerical results demonstrating and validating the theory are presented. Finally, in the remaining sections, an application to a practical problem arising in data assimilation is described, and the conclusions are summarized.

2. GN method.

We begin by introducing the GN method and reviewing briefly some results on the convergence of the method. We then define the truncated and perturbed approximate GN methods that will be examined in subsequent sections.

2.1. Statement of the algorithm.

We consider the NLSP defined in (1), where we assume that

A0. $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ is a nonlinear twice continuously Fréchet differentiable function.

We denote the Jacobian of the function f by $J(x) \equiv f'(x)$. The gradient and Hessian of $\phi(x)$ are then given by

$$\nabla\phi(x) = J(x)^T f(x), \tag{2}$$

$$\nabla^2\phi(x) = J(x)^T J(x) + Q(x), \tag{3}$$

where $Q(x)$ denotes the second-order terms

$$Q(x) = \sum_{i=1}^{m} f_i(x) \nabla^2 f_i(x). \tag{4}$$

The following additional assumptions are made in order to establish the theory:

A1. There exists $x^* \in \mathbb{R}^n$ such that $J(x^*)^T f(x^*) = 0$;

A2. The Jacobian matrix $J(x^*)$ at x^* has full rank n.

Finding the stationary points of ϕ is equivalent to solving the gradient equation

$$F(x) \equiv \nabla \phi(x) = J(x)^T f(x) = 0. \tag{5}$$

Techniques for treating the NLSP can thus be derived from methods for solving this nonlinear algebraic system.

A common method for solving nonlinear equations of form (5) and hence for solving the NLSP (1) is Newton's method [20, section 10.2]. This method requires the full Hessian matrix (3) of function ϕ. For many large scale problems, the second-order terms $Q(x)$ of the Hessian are, however, impracticable to calculate and, in order to make the procedure more efficient, Newton's method is approximated by ignoring these terms. The resulting iterative method is known as the GN algorithm [20, section 8.5] and is defined as follows.

ALGORITHM GN ALGORITHM.

Step 0. Choose an initial $x_0 \in \mathbb{R}^n$.

Step 1. Repeat until convergence:

Step 1.1. Solve $J(x_k)^T J(x_k) s_k = -J(x_k)^T f(x_k)$.

Step 1.2. Set $x_{k+1} = x_k + s_k$.

Remarks. We note that at each iteration, Step 1.1 of the method corresponds to solving the linearized least squares problem

$$\min_s \frac{1}{2} \left\| J(x_k) s + f(x_k) \right\|_2^2. \tag{6}$$

We note also that the GN method can be written as a fixed-point iteration of the form

$$x_{k+1} = G(x_k), \tag{7}$$

where $G(x) \equiv x - J^+(x) f(x)$ and $J^+(x) \equiv (J(x)^T J(x))^{-1} J(x)^T$ denotes the Moore–Penrose pseudoinverse of $J(x)$.

2.2. Convergence of the exact GN method. Sufficient conditions for the convergence of the GN method are known in the case where the normal equations for the linearized least squares problem (6) are solved *exactly* in Step 1.1 at each iteration. We now recall some existing results.

We introduce the notation $\rho(A)$ to indicate the spectral radius of an $n \times n$ matrix A, and we define

$$\varrho = \rho\left(\left(J(x^*)^T J(x^*) \right)^{-1} Q(x^*) \right). \tag{8}$$

The following theorem on local convergence of the GN method then holds.

THEOREM 1 (Ortega and Rheinboldt [20, Theorem 10.1.3]). *Let assumptions* A0, A1, *and* A2 *hold. If* $\varrho < 1$, *then the GN iteration converges locally to* x^*; *that is, there*

exists $\varepsilon > 0$ *such that the sequence* $\{x_k\}$ *generated by the GN algorithm converges to* x^* *for all* $x_0 \in \mathcal{D} \equiv \{x \mid \|x - x^*\|_2 < \varepsilon\}$.

Theorem 1 has a geometrical interpretation as described in [23] (see also [2, section 9.2.2]). We denote by $\mathcal{S}$ the surface in $\mathbb{R}^m$ given by the parametric representation $y = f(x)$, $x \in \mathbb{R}^n$, and we let M be the point on $\mathcal{S}$ with coordinates $f(x^*)$, taking O as the origin of the coordinate system. The vector OM is orthogonal to the plane tangent to the surface $\mathcal{S}$ through M.

THEOREM 2 (Wedin [23]). *Suppose that the assumptions of Theorem* 1 *hold and that* $f(x^*)$ *is nonzero. Then*

$$\varrho = \|f(x^*)\|_2 \chi, \tag{9}$$

where χ *is the maximal principal curvature of the surface* $\mathcal{S}$ *at point* M *with respect to the normal direction* $w^* = f(x^*)/\|f(x^*)\|_2$.

In the *zero residual* case, where $f(x^*) = 0$, the relation (9) continues to hold. In this case the origin O lies on the surface $\mathcal{S}$ and χ denotes the maximal principal curvature of $\mathcal{S}$ with respect to the direction normal to the tangent surface at O. Since we then have $Q(x^*) = 0$ and hence $\varrho = 0$, the result still holds.

For the GN method to converge it is therefore sufficient for the maximal principal curvature χ of the surface $\mathcal{S}$ at the point $f(x^*)$ to satisfy $1/\chi > \|f(x^*)\|_2$. This condition holds if and only if $\nabla^2\phi(x^*)$ is positive definite at x^* and ensures that x^* is a local minimizer of the objective function ϕ [2, section 9.1.2]. The relation (9) implies that the convergence condition of Theorem 1 is invariant under transformation of the NLSP by a local diffeomorphism, since the quantity $\|f(x^*)\|_2 \chi$ has this property [23].

The proofs of these results depend on theory for stationary fixed point iteration processes [20]. The theory ensures local convergence at a linear rate. Additional, but more restrictive, conditions for local convergence are given in [9]. Conditions giving higher-order rates of convergence can be deduced from this theory. The GN method can also be treated as an IN method [21, 8, 3, 4]. Results of these types will be discussed further in sections 4 and 5.

We remark that the GN method may not be locally convergent in some cases [9]. Nevertheless, approximate GN methods are used widely in practice for solving the very large NLSP problems arising in data assimilation. The approximations are designed to make the algorithm more computationally efficient. Our aim here is to investigate the effects of these approximations on the convergence of the GN algorithm and to establish conditions for the local convergence of the approximate methods, given that conditions hold for the exact GN method to be locally convergent.

3. Approximate GN algorithms. A serious difficulty associated with the use of the GN method in large scale applications, such as data assimilation, is that the linearized least squares problem (6) is computationally too expensive to solve exactly in Step 1.1 of the algorithm at each iteration. The dimensions of the normal matrix equations to be solved in Step 1.1 are often so great that the system coefficients cannot be stored in core memory, even in factored form. Therefore, in order to solve the full nonlinear problem efficiently, in real forecasting time, approximations must be made within the GN procedure.

Two types of approximation are commonly applied. First, the linearized least squares problem (6) is solved only approximately by an inner iteration method that is truncated before full accuracy is reached. We refer to this approximate algorithm as the TGN method. Second, the linearized least squares problem in Step 1.1 is replaced by an approximate, simplified or perturbed, linear problem that can be solved more

efficiently in the inner loop. We refer to this algorithm as the PGN method. Here we examine both of these approximate GN methods and also the combined truncated perturbed GN (TPGN) method, where both approximations are applied. In the next subsections we define these procedures explicitly, and in sections 4 and 5 we analyze the convergence of the approximate methods.

3.1. TGN method. At each outer iteration k of the GN method, we solve the normal equations

$$J(x_k)^T J(x_k) s = -J(x_k)^T f(x_k) \tag{10}$$

for the linearized least squares problem (6) using an iterative procedure. Intuitively, when x_k is far from x^* and the function f is nonlinear, it is not worth solving (10) to high accuracy. A natural stopping criterion for the iterative process is where the relative residual satisfies

$$\left\| J(x_k)^T J(x_k) s_k + J(x_k)^T f(x_k) \right\|_2 / \left\| J(x_k)^T f(x_k) \right\|_2 \leq \beta_k. \tag{11}$$

Here s_k denotes the current estimate of the solution of (10), and β_k is a specified tolerance. For this reason we define the TGN algorithm as follows.

Algorithm TGN algorithm.

Step 0. Choose an initial $x_0 \in \mathbb{R}^n$.

Step 1. Repeat until convergence:

Step 1.1. Find s_k such that
$(J(x_k)^T J(x_k)) s_k = -J(x_k)^T f(x_k) + r_k$
with $\|r_k\|_2 \leq \beta_k \left\| J(x_k)^T f(x_k) \right\|_2$.

Step 1.2. Update $x_{k+1} = x_k + s_k$.

The tolerances β_k, $k = 0, 1, 2 \ldots$, must be chosen to ensure convergence of the procedure to the optimal x^* of the NLSP (1). Conditions guaranteeing convergence of the TGN method are presented in sections 4 and 5.

3.2. PGN method. For some applications it is desirable to apply a PGN method in which the true Jacobian J is replaced by an approximation $\tilde{J}$; this is practical, for example, in cases where a perturbed Jacobian is much easier or computationally less expensive to calculate. We therefore define the PGN method as follows.

Algorithm PGN algorithm.

Step 0. Choose an initial $x_0 \in \mathbb{R}^n$.

Step 1. Repeat until convergence:

Step 1.1. Solve $\tilde{J}(x_k)^T \tilde{J}(x_k) s_k = -\tilde{J}(x_k)^T f(x_k)$.

Step 1.2. Set $x_{k+1} = x_k + s_k$.

We emphasize that in Step 1.1 of the PGN algorithm only the Jacobian is approximated and not the nonlinear function $f(x_k)$. The approximate Jacobian, $\tilde{J}(x)$, is assumed to be continuously Fréchet differentiable.

In applications to data assimilation the approximate Jacobian is derived from an approximate linearization of the discrete nonlinear model equations, and hence the perturbed Jacobian approximates the same underlying dynamical system as the exact Jacobian and has similar properties. The assumptions made here and in sections 4.4–4.6 on the perturbed Jacobian are therefore regarded as reasonable. The derivation of the perturbed Jacobian for a practical data assimilation problem is shown in section 7.

In order to interpret the PGN iteration, it is convenient to define the function

$$\tilde{F}(x) = \tilde{J}(x)^T f(x) \tag{12}$$

and to write its first derivative in the form

$$\tilde{F}'(x) = \tilde{J}(x)^T J(x) + \tilde{Q}(x), \tag{13}$$

where $J(x)$ is the Jacobian of the function $f(x)$ and $\tilde{Q}(x)$ represents second-order terms arising from the derivative of $\tilde{J}(x)$. Then the PGN algorithm can be considered as an iterative method for finding a solution $\tilde{x}^*$ to the nonlinear equation

$$\tilde{F}(x) = 0. \tag{14}$$

We remark that, just as the GN method can be regarded as an IN method for solving the gradient equation (5), the PGN method can be treated as an IN method for solving the perturbed gradient equation (14). In the PGN method, the second-order term in the derivative $\tilde{F}'$ is ignored and the first-order term is now approximated, allowing the iteration to be written as a sequence of linear least squares problems.

For the zero residual NLSP, where $f(x^*) = 0$, the solution x^* of the problem satisfies (14) and so the fixed point $\tilde{x}^* = x^*$ of the PGN procedure is also a fixed point of the exact GN iteration. Similarly, if $f(x^*)$ lies in the null space of $\tilde{J}(x^*)$, then (14) is satisfied by x^* and the fixed point of the PGN method is again a fixed point of the exact GN method. In general, the fixed point of the PGN method will not be the same as that of the GN algorithm. We might expect, however, that if $\tilde{J}$ is close to J, then the solution $\tilde{x}^*$ of (14) will be close to the solution x^* of the true gradient equation (5).

In sections 4 and 5 we give conditions for the PGN method to converge locally, and in section 4 we also examine the distance between the fixed points of the two algorithms.

3.3. TPGN method. In the PGN method, we solve the normal equations in Step 1.1 of the algorithm at each outer iteration k by applying an iterative method to the perturbed linear least squares problem

$$\min_s \frac{1}{2} \left\| \tilde{J}(x_k)s + f(x_k) \right\|_2^2. \tag{15}$$

To improve the efficiency of the PGN procedure, the iteration is truncated before full accuracy is reached. The iterations are halted where the relative residual satisfies

$$\left\| \tilde{J}(x_k)^T \tilde{J}(x_k)s_k + \tilde{J}(x_k)^T f(x_k) \right\|_2 / \left\| \tilde{J}(x_k)^T f(x_k) \right\|_2 \leq \beta_k. \tag{16}$$

Here s_k is the current estimate of the solution of (15), and β_k is a specified tolerance. This procedure is referred to as the TPGN method and is defined as follows.

ALGORITHM TPGN ALGORITHM.

Step 0. Choose an initial $x_0 \in \mathbb{R}^n$.

Step 1. Repeat until convergence:

Step 1.1. Find s_k such that

$$\tilde{J}(x_k)^T \tilde{J}(x_k)s_k = -\tilde{J}(x_k)^T f(x_k) + r_k$$

$$\text{with } \|r_k\|_2 \leq \beta_k \left\| \tilde{J}(x_k)^T f(x_k) \right\|_2.$$

Step 1.2. Update $x_{k+1} = x_k + s_k$.

The tolerances β_k, $k = 0, 1, 2 \ldots$, must be chosen to ensure convergence of the procedure to the optimal $\tilde{x}^*$ of the perturbed gradient equation (14). Conditions guaranteeing local convergence of the TPGN method are presented in sections 4 and 5.

4. Convergence of approximate GN methods I. We now derive sufficient conditions for the convergence of the truncated and perturbed GN methods. The theory is based on two different approaches. In this section we present results based on theory for IN methods found in [8] and [3]. In the subsequent section we extend the arguments of [9] for exact GN methods to the approximate truncated and perturbed methods. The aim is to establish conditions for the convergence of the approximate methods, given that conditions hold for the exact method to converge.

We begin by introducing the theory for IN methods. This theory is applied to the exact GN method to obtain a new convergence condition. Criteria for the convergence of the truncated and perturbed methods are then derived using these results.

4.1. IN methods. The IN method for solving the NLSP problem (1), as defined in [8], is given as follows.

ALGORITHM IN ALGORITHM.

Step 0. Choose an initial $x_0 \in \mathbb{R}^n$.

Step 1. Repeat until convergence:

Step 1.1. Solve $\nabla^2\phi(x_k)s_k = -\nabla\phi(x_k) + \tilde{r}_k$.

Step 1.2. Set $x_{k+1} = x_k + s_k$.

In Step 1.1 the residual errors $\tilde{r}_k$ measure the amount by which the calculated solution s_k fails to satisfy the exact Newton method at each iteration. It is assumed that the relative sizes of these residuals are bounded by a nonnegative forcing sequence $\{\eta_k\}$ such that for each iteration

$$\frac{\|\tilde{r}_k\|_2}{\|\nabla\phi(x_k)\|_2} \leq \eta_k. \tag{17}$$

Conditions for the convergence of the IN algorithm are established in the following theorem.

THEOREM 3 (Dembo, Eisenstat, and Steihaug [8]). *Let assumptions* A0, A1, *and* A2 *hold, and let* $\nabla^2\phi(x^*)$ *be nonsingular. Assume* $0 \leq \eta_k \leq \hat{\eta} < t < 1$. *Then there exists* $\varepsilon > 0$ *such that, if* $\|x_0 - x^*\|_2 \leq \varepsilon$, *the sequence of IN iterates* $\{x_k\}$ *satisfying* (17) *converges to* x^*. *Moreover, the convergence is linear in the sense that*

$$\| x_{k+1} - x^* \|_* \leq t \| x_k - x^* \|_*, \tag{18}$$

where $\| y \|_* = \left\|\nabla^2\phi(x^*)y\right\|_2$.

In [3] Theorem 3 is applied to obtain more general results in which the Jacobian and Hessian matrices are perturbed on each iteration of the Newton method. Here we adopt similar techniques to derive results for the approximate GN methods based on theory for the IN methods.

4.2. GN as an IN method. We first establish novel sufficient conditions for the exact GN method to converge by treating it as an IN method.

THEOREM 4. *Let assumptions* A0, A1, *and* A2 *hold, and let* $\nabla^2\phi(x^*)$ *be nonsingular. Assume* $0 \leq \hat{\eta} < 1$. *Then there exists* $\varepsilon > 0$ *such that, if* $\|x_0 - x^*\|_2 \leq \varepsilon$ *and if*

$$\left\|Q(x_k)(J(x_k)^T J(x_k))^{-1}\right\|_2 \leq \eta_k \leq \hat{\eta} \quad \text{for} \quad k = 0, 1, \ldots, \tag{19}$$

the sequence of GN iterates $\{x_k\}$ *converges to* x^*.

Proof of Theorem 4. We can write the GN method as an IN method by setting

$$\begin{aligned} \tilde{r}_k &= \nabla\phi(x_k) - \nabla^2\phi(x_k)(J(x_k)^T J(x_k))^{-1}\nabla\phi(x_k) \\ &= (I - \nabla^2\phi(x_k)(J(x_k)^T J(x_k))^{-1})\nabla\phi(x_k). \end{aligned} \tag{20}$$

Then, using (3), we have

$$\begin{aligned} \|\tilde{r}_k\|_2 &= \left\|(I - \nabla^2\phi(x_k)(J(x_k)^T J(x_k))^{-1})\nabla\phi(x_k)\right\|_2 \\ &\leq \left\|Q(x_k)(J(x_k)^T J(x_k))^{-1}\right\|_2 \|\nabla\phi(x_k)\|_2 . \end{aligned} \tag{21}$$

By Theorem 3, a sufficient condition for local convergence is therefore

$$\left\|Q(x_k)(J(x_k)^T J(x_k))^{-1}\right\|_2 \leq \eta_k \leq \hat{\eta}, \quad k = 0, 1, \ldots . \qquad \square \tag{22}$$

The convergence condition derived in this theorem is more restrictive than that obtained in Theorem 1, which requires a bound only on the spectral radius of the matrix $Q(x)(J(x)^T J(x))^{-1}$ at the fixed point $x = x^*$ rather than on its norm at each iterate x_k. The technique used in the proof of Theorem 4 is, however, more readily extended to the case of the approximate GN iterations and enables qualitative information on the conditions needed for convergence to be established. This approach also provides a practical test of convergence for the approximate methods (see [17]).

4.3. Convergence of the TGN method (I). We now give a theorem that provides sufficient conditions for the convergence of the TGN method. It is assumed that the residuals in the TGN method are bounded such that

$$\|r_k\|_2 \leq \beta_k \|\nabla\phi(x_k)\|_2 , \tag{23}$$

where $\{\beta_k\}$ is a nonnegative forcing sequence. The theorem is established by considering the algorithm as an IN method, as in the proof of Theorem 4.

THEOREM 5. *Let assumptions* A0, A1, *and* A2 *hold, and let* $\nabla^2\phi(x^*)$ *be nonsingular. Assume that* $0 \leq \hat{\beta} < 1$, *and select* β_k, $k = 0, 1, \ldots$, *such that*

$$0 \leq \beta_k \leq \frac{\hat{\beta} - \left\|Q(x_k)(J(x_k)^T J(x_k))^{-1}\right\|_2}{1 + \left\|Q(x_k)(J(x_k)^T J(x_k))^{-1}\right\|_2}, \quad k = 0, 1, \ldots . \tag{24}$$

Then there exists $\varepsilon > 0$ *such that, if* $\|x_0 - x^*\|_2 \leq \varepsilon$, *the sequence of TGN iterates* $\{x_k\}$ *satisfying* (23) *converges to* x^*.

Proof of Theorem 5. We can write the TGN method as an IN method by setting

$$\tilde{r}_k = \nabla\phi(x_k) - \nabla^2\phi(x_k)(J(x_k)^T J(x_k))^{-1}\nabla\phi(x_k) + \nabla^2\phi(x_k)(J(x_k)^T J(x_k))^{-1} r_k. \tag{25}$$

Then we have

$$\begin{aligned} \|\tilde{r}_k\|_2 &\leq \left\|Q(x_k)(J(x_k)^T J(x_k))^{-1}\right\|_2 \|\nabla\phi(x_k)\|_2 + \left\|I + Q(x_k)(J(x_k)^T J(x_k))^{-1}\right\|_2 \|r_k\|_2 \\ &\leq \left(\left\|Q(x_k)(J(x_k)^T J(x_k))^{-1}\right\|_2 + \beta_k(1 + \left\|Q(x_k)(J(x_k)^T J(x_k))^{-1}\right\|_2)\right) \|\nabla\phi(x_k)\|_2 \\ &\leq \left(\left\|Q(x_k)(J(x_k)^T J(x_k))^{-1}\right\|_2 + (\hat{\beta} - \left\|Q(x_k)(J(x_k)^T J(x_k))^{-1}\right\|_2)\right) \|\nabla\phi(x_k)\|_2 \\ &\leq \hat{\beta} \|\nabla\phi(x_k)\|_2 . \end{aligned} \tag{26}$$

Local convergence then follows from Theorem 3. $\square$

Since $\beta_k \geq 0$ is necessary, we require that $\left\|Q(x_k)(J(x_k)^T J(x_k))^{-1}\right\|_2 \leq \hat{\beta} < 1$. This is just the sufficient condition given by Theorem 4 for the exact GN method to converge.

We remark also that the more highly nonlinear the problem is, the larger the norm $\left\|Q(x_k)(J(x_k)^T J(x_k))^{-1}\right\|_2$ will be and hence the smaller the limit on β_k will be. The inner iteration of the TGN method must then be solved more accurately to ensure convergence of the algorithm.

4.4. Convergence of the PGN method (I). Next we present sufficient conditions for the PGN method to converge. The theorem is established by considering the PGN method as an IN method for solving the perturbed gradient equation (14). We make the assumptions:

A1′. There exists $\tilde{x}^* \in \mathbb{R}^n$ such that $\tilde{F}(\tilde{x}^*) \equiv \tilde{J}(\tilde{x}^*)^T f(\tilde{x}^*) = 0$;

A2′. The matrix $\tilde{J}(\tilde{x}^*)$ at $\tilde{x}^*$ has full rank n.

We then obtain the theorem.

THEOREM 6. *Let assumptions* A0, A1′, *and* A2′ *hold, and let* $\tilde{F}'(\tilde{x}^*) \equiv \tilde{J}(\tilde{x}^*)^T J(\tilde{x}^*) + \tilde{Q}(\tilde{x}^*)$ *be nonsingular. Assume* $0 \le \hat{\eta} < 1$. *Then there exists* $\varepsilon > 0$ *such that, if* $\|x_0 - \tilde{x}^*\|_2 \le \varepsilon$ *and if*

$$\left\| I - (\tilde{J}(x_k)^T J(x_k) + \tilde{Q}(x_k))(\tilde{J}(x_k)^T \tilde{J}(x_k))^{-1} \right\|_2 \le \eta_k \le \hat{\eta}, \quad k = 0, 1, \ldots, \tag{27}$$

the sequence of perturbed GN iterates $\{x_k\}$ *converges to* $\tilde{x}^*$.

Proof of Theorem 6. We can write the PGN method as an IN method by setting

$$\tilde{r}_k = \tilde{J}(x_k)^T f(x_k) - (\tilde{J}(x_k)^T J(x_k) + \tilde{Q}(x_k))(\tilde{J}(x_k)^T \tilde{J}(x_k))^{-1} \tilde{J}(x_k)^T f(x_k) \tag{28}$$

$$= (I - (\tilde{J}(x_k)^T J(x_k) + \tilde{Q}(x_k))(\tilde{J}(x_k)^T \tilde{J}(x_k))^{-1}) \tilde{J}(x_k)^T f(x_k). \tag{29}$$

Then, provided the condition (27) holds, we have

$$\|\tilde{r}_k\|_2 \le \hat{\eta} \left\| \tilde{J}(x_k)^T f(x_k) \right\|_2, \tag{30}$$

and by Theorem 3 local convergence is guaranteed. ∎

The theorem gives explicit conditions on the perturbed Jacobian $\tilde{J}$ that are sufficient to guarantee the convergence of the PGN method. The requirement is that $\tilde{J}(x)^T \tilde{J}(x)$ should be a good approximation to the derivative $\tilde{F}'(x) = \tilde{J}(x)^T J(x) + \tilde{Q}(x)$ of the perturbed gradient equation (14).

4.5. Fixed point of the PGN method. We now consider how close the solution $\tilde{x}^*$ of the perturbed gradient equation (14) is to the solution x^* of the original NLSP. To answer this question we treat the GN method as a stationary fixed-point iteration of the form (7).

We assume that the GN iteration converges locally to x^* for all x_0 in an open convex set $\mathcal{D}$ containing x^* (defined as in Theorem 1) and that $G(x)$ satisfies

$$\|G(x) - G(x^*)\|_2 \le \nu \|x - x^*\|_2 \quad \forall\, x \in \mathcal{D} \ \text{ with } \ \nu < 1, \tag{31}$$

where $G(x)$ is as given in section 2.1. Then we have the following theorem, which bounds the distance between the solutions of the exact and perturbed iterations.

THEOREM 7. *Let assumptions* A0, A1, A2, A1′, *and* A2′ *hold, and assume* $\varrho < 1$. *Let* (31) *be satisfied. Also let* $\tilde{x}^* \in \mathcal{D}$, *and assume* $J(\tilde{x}^*)$ *is of full rank. Then*

$$\|\tilde{x}^* - x^*\|_2 \le \frac{1}{1-\nu} \left\| (\tilde{J}^+(\tilde{x}^*) - J^+(\tilde{x}^*)) f(\tilde{x}^*) \right\|_2. \tag{32}$$

Proof of Theorem 7. We define $\tilde{G}(x) = x - \tilde{J}^+(x) f(x)$. Then $\tilde{x}^* = \tilde{G}(\tilde{x}^*)$, and we have

$$\begin{aligned} \|\tilde{x}^* - x^*\|_2 &= \left\| \tilde{G}(\tilde{x}^*) - G(x^*) \right\|_2 \\ &\le \left\| \tilde{G}(\tilde{x}^*) - G(\tilde{x}^*) \right\|_2 + \|G(\tilde{x}^*) - G(x^*)\|_2 \\ &\le \nu \|\tilde{x}^* - x^*\|_2 + \left\| \tilde{G}(\tilde{x}^*) - G(\tilde{x}^*) \right\|_2. \end{aligned}$$

Hence, we obtain

$$\begin{aligned}\|\tilde{x}^* - x^*\|_2 &\le \frac{1}{1-\nu}\left\|\tilde{G}(\tilde{x}^*) - G(\tilde{x}^*)\right\|_2 \\ &\le \frac{1}{1-\nu}\left\|(\tilde{J}^+(\tilde{x}^*) - J^+(\tilde{x}^*))f(\tilde{x}^*)\right\|_2. \qquad \square\end{aligned}$$

The theorem shows that the distance between x^* and $\tilde{x}^*$ is bounded in terms of the distance between the pseudoinverses of $\tilde{J}$ and J at $\tilde{x}^*$ and will be small if these are close together. The theorem also implies, from (14), that the bound given in (32) equals $\|J^+(\tilde{x}^*)f(\tilde{x}^*)\|_2/(1-\nu)$, which is proportional to the residual in the true gradient equation (5) evaluated at the solution $\tilde{x}^*$ of the perturbed gradient equation (14).

A different approach to the convergence of the perturbed fixed point iteration can be found in [20, Theorem 12.2.5]. This approach shows essentially that if the GN method converges, then the PGN iterates eventually lie in a small region around x^* of radius $\delta/(1-\nu)$, where δ bounds the distance $\|\tilde{G}(x) - G(x)\|_2$ over all $x \in \mathcal{D}$. This theory does not establish convergence of the perturbed method, but the theory for the distance between the fixed points of the GN and PGN methods presented here is consistent with these results.

4.6. Convergence of the TPGN method (I). We now examine the convergence of the approximate GN method where the Jacobian is perturbed and the inner linear least squares problem is not solved exactly. The residuals in the inner normal equations at each outer iteration are assumed to be bounded such that

$$\|r_k\|_2 \le \beta_k \left\|\tilde{J}(x_k)^T f(x_k)\right\|_2, \tag{33}$$

where $\{\beta_k\}$ is a nonnegative forcing sequence. Sufficient conditions for the convergence of this TPGN method are given by the next theorem.

THEOREM 8. *Let assumptions* A0, A1′, *and* A2′ *hold, and let* $\tilde{F}'(\tilde{x}^*) \equiv \tilde{J}(\tilde{x}^*)^T J(\tilde{x}^*) + \tilde{Q}(\tilde{x}^*)$ *be nonsingular. Assume that* $0 \le \hat{\beta} < 1$, *and select* β_k, $k = 0, 1, \ldots,$ *such that*

$$0 \le \beta_k \le \frac{\hat{\beta} - \left\|I - (\tilde{J}(x_k)^T J(x_k) + \tilde{Q}(x_k))(\tilde{J}(x_k)^T \tilde{J}(x_k))^{-1}\right\|_2}{\left\|(\tilde{J}(x_k)^T J(x_k) + \tilde{Q}(x_k))(\tilde{J}(x_k)^T \tilde{J}(x_k))^{-1}\right\|_2}. \tag{34}$$

Then there exists $\varepsilon > 0$ *such that, if* $\|x_0 - \tilde{x}^*\|_2 \le \varepsilon$, *the sequence of PGN iterates* $\{x_k\}$ *satisfying* (33) *converges to* $\tilde{x}^*$.

Proof of Theorem 8. We can write TPGN in the same form as IN by setting

$$\begin{aligned}\tilde{r}_k = {} & (I - (\tilde{J}(x_k)^T J(x_k) + \tilde{Q}(x_k))(\tilde{J}(x_k)^T \tilde{J}(x_k))^{-1})\tilde{J}(x_k)^T f(x_k) \\ & + (\tilde{J}(x_k)^T J(x_k) + \tilde{Q}(x_k))(\tilde{J}(x_k)^T \tilde{J}(x_k))^{-1} r_k.\end{aligned} \tag{35}$$

Then, provided the condition (33) holds, we have

$$\|\tilde{r}_k\|_2 \le \hat{\beta}\left\|\tilde{J}(x_k)^T f(x_k)\right\|_2, \tag{36}$$

and by Theorem 3 local convergence is guaranteed. $\square$

We remark that in order to ensure $\beta_k \geq 0$ we also require that

$$\left\| I - (\tilde{J}(x_k)^T J(x_k) + \tilde{Q}(x_k))(\tilde{J}(x_k)^T \tilde{J}(x_k))^{-1} \right\|_2 \leq \hat{\beta} < 1,$$

which is simply the sufficient condition found in Theorem 6 for the PGN method to converge.

4.7. Summary. In this section we have established theory ensuring local linear convergence of the GN, the TGN, the PGN, and the TPGN methods based on the theory of [8] for IN methods. Numerical examples illustrating the results for the three approximate GN methods are shown in section 6, and a practical application to data assimilation is presented in section 7. In the next section we derive additional convergence conditions for these methods based on the theory of [9] for exact GN methods.

5. Convergence of approximate GN methods II. We now derive conditions for the convergence of the approximate GN methods by extending the results of [9] for the exact GN method. These results are more restrictive than those given in section 4 but provide more precise estimates of the rates of convergence of the methods. Conditions for linear, superlinear, and quadratic convergence are established.

5.1. Sufficient conditions for the exact GN method. We begin by recalling the sufficient conditions of [9] for local convergence of the GN iterates to a stationary point x^* of the NLSP.

THEOREM 9 (Dennis and Schnabel [9, Theorem 10.2.1]). *Let assumptions* A0, A1, *and* A2 *hold, and let λ be the smallest eigenvalue of the matrix $J(x^*)^T J(x^*)$. Suppose that there exists an open convex set $\mathcal{D}$ containing x^* such that*

(i) *$J(x)$ is Lipschitz continuous in $\mathcal{D}$ with a Lipschitz constant equal to γ;*
(ii) *$\|J(x)\|_2 \leq \alpha$ for all $x \in \mathcal{D}$;*
(iii) *there exists $\sigma \geq 0$ such that $\left\|(J(x) - J(x^*))^T f(x^*)\right\|_2 \leq \sigma \|x - x^*\|_2$ for all $x \in \mathcal{D}$;*
(iv) *$\sigma < \lambda$.*

Let c be such that $1 < c < \lambda/\sigma$. Then there exists $\varepsilon > 0$ such that, if $\|x_0 - x^*\|_2 < \varepsilon$, the iterates $\{x_k\}$ generated by the GN algorithm converge to x^*. Additionally, the following inequality holds:

$$\|x_{k+1} - x^*\|_2 \leq \frac{c\sigma}{\lambda} \|x_k - x^*\|_2 + \frac{c\alpha\gamma}{2\lambda} \|x_k - x^*\|_2^2 . \tag{37}$$

The constant σ may be regarded as an approximation to the norm of the second-order terms $\|Q(x^*)\|_2$ and is a combined measure of the nonlinearity of the problem and the size of the residual [9, section 10.2]. The theorem shows that the convergence of the GN method is quadratic in the case $\sigma = 0$. This holds, for example, for the zero-residual problem where $f(x^*) = 0$.

The sufficient conditions given by Theorem 9 for the local convergence of the GN method are more restrictive than those given in Theorem 1. We demonstrate this as follows.

THEOREM 10. *If the assumptions of Theorem* 9 *hold, then $\varrho < 1$.*

Proof of Theorem 10. By Taylor expansion of the map $x \mapsto J(x)^T f(x^*)$ with respect to x, we find

$$J(x)^T f(x^*) = Q(x^*)(x - x^*) + \|x - x^*\|_2 \, \Theta(x - x^*), \tag{38}$$

with $\lim_{h\to 0}\Theta(h) = 0$. We denote $f(x^*)$, $J(x^*)$, and $Q(x^*)$ by f^*, J^*, and Q^*, respectively. Then multiplying (38) by $(J^{*T}J^*)^{-1}$ on the left yields

$$(J^{*T}J^*)^{-1}J(x)^T f^* = (J^{*T}J^*)^{-1}Q^*(x-x^*) + \|x-x^*\|_2\,\Theta_1(x-x^*), \tag{39}$$

with $\lim_{h\to 0}\Theta_1(h) = 0$. We let v be the right singular vector associated with the largest singular value of $(J^{*T}J^*)^{-1}Q^*$ and let $x_\epsilon = x^* + \epsilon v$ for $\epsilon > 0$. Substituting x_ϵ for x in (39) and rearranging the terms of the equality then gives us

$$\epsilon(J^{*T}J^*)^{-1}Q^*v = (J^{*T}J^*)^{-1}J(x_\epsilon)^T f^* - \epsilon\Theta_1(\epsilon v). \tag{40}$$

By the assumptions of Theorem 9, we have $\left\|J(x_\epsilon)^T f^*\right\|_2 \le \sigma\epsilon$ for ϵ sufficiently small, and therefore

$$\left\|(J^{*T}J^*)^{-1}J(x_\epsilon)^T f^*\right\|_2 \le \left\|(J^{*T}J^*)^{-1}\right\|_2 \sigma\epsilon = \epsilon\sigma/\lambda. \tag{41}$$

Taking norms in (40) and letting ϵ tend to 0 then yields

$$\left\|(J^{*T}J^*)^{-1}Q^*\right\|_2 \le \sigma/\lambda. \tag{42}$$

Since

$$\varrho \le \left\|(J^{*T}J^*)^{-1}Q^*\right\|_2, \tag{43}$$

we obtain $\varrho \le \sigma/\lambda$. Therefore, if $\sigma < \lambda$, then $\varrho < 1$. □

The conditions of Theorem 9 ensure that the conditions of Theorem 1 hold and that the exact GN method converges, but the conditions of Theorem 1 are weaker than those of Theorem 9. Since the quantity $\sigma > \|Q(x^*)\|_2$ can be made arbitrarily close to $\|Q(x^*)\|_2$ in a sufficiently small neighborhood of x^*, the condition $\sigma < \lambda$ can be achieved only if $\|Q(x^*)\|_2 \left\|(J(x^*)^T J(x^*)^T)^{-1}\right\|_2 < 1$, which is a stronger requirement than that of Theorem 1 for convergence (see [12]).

We now extend the theory of Theorem 9 to the approximate GN methods. The results are not as general as those of section 4 but allow the rates of convergence of the methods to be determined.

5.2. Convergence of the TGN method (II). By an extension of Theorem 9, we now establish alternative conditions for the TGN method to converge. We assume, as previously, that the residuals in the TGN method are bounded such that

$$\|r_k\|_2 \le \beta_k \left\|J(x_k)^T f(x_k)\right\|_2, \tag{44}$$

where $\{\beta_k\}$ is a nonnegative forcing sequence.

THEOREM 11. *Let the conditions of Theorem* 9 *hold, and let* c *be such that* $1 < c < \lambda/\sigma$. *Select* β_k, $k = 0, 1, \ldots$, *to satisfy*

$$0 \le \beta_k \le \hat{\beta} < \frac{\lambda - c\sigma}{c(\sigma + \alpha^2)}, \quad k = 0, 1, \ldots. \tag{45}$$

Then there exists $\varepsilon > 0$ *such that, if* $\|x_0 - x^*\|_2 < \varepsilon$, *the sequence of TGN iterates* $\{x_k\}$ *satisfying* (44) *converges to* x^*. *Additionally, the following inequality holds:*

$$\|x_{k+1} - x^*\|_2 \le \frac{c}{\lambda}(\sigma + \beta_k(\sigma + \alpha^2))\,\|x_k - x^*\|_2 + C\,\|x_k - x^*\|_2^2, \tag{46}$$

where $C = \frac{c\alpha\gamma}{2\lambda}(1 + \hat{\beta})$.

Proof of Theorem 11. The proof is by induction. Let $k = 0$, and denote by J_0, f_0, J^*, and f^* the quantities $J(x_0)$, $f(x_0)$, $J(x^*)$, and $f(x^*)$, respectively. From the proof of Theorem 9 (see [9, Theorem 10.2.1]), there exists a positive quantity ε_1 such that, if $\|x_0 - x^*\|_2 < \varepsilon_1$, then $x_0 \in \mathcal{D}$, $J_0^T J_0$ is nonsingular, $\left\|(J_0^T J_0)^{-1}\right\|_2 \leq c/\lambda$, and

$$\left\|x_0 - (J_0^T J_0)^{-1} J_0^T f_0 - x^*\right\|_2 \leq \frac{c\sigma}{\lambda} \|x_0 - x^*\|_2 + \frac{c\alpha\gamma}{2\lambda} \|x_0 - x^*\|_2^2 . \tag{47}$$

Let

$$\varepsilon = \min\left\{\varepsilon_1, \frac{\lambda - c(\sigma + \hat{\beta}(\sigma + \alpha^2))}{c\alpha\gamma(1 + \hat{\beta})}\right\}, \tag{48}$$

where $\lambda - c(\sigma + \hat{\beta}(\sigma + \alpha^2)) > 0$ by (45).

We start from

$$\left\|J_0^T f_0\right\|_2 = \left\|J_0^T f^* + J_0^T (J_0(x_0 - x^*) + f_0 - f^*) - J_0^T J_0 (x_0 - x^*)\right\|_2 \tag{49}$$

and bound successively each term in the norm. From the definitions of σ and α in Theorem 9, we have $\left\|J_0^T f^*\right\|_2 \leq \sigma \|x_0 - x^*\|_2$ and $\left\|J_0^T J_0 (x_0 - x^*)\right\|_2 \leq \alpha^2 \|x_0 - x^*\|_2$. From [9, Lemma 4.1.12] and the Lipschitz continuity of J_0, we also have

$$\|J_0(x_0 - x^*) + f^* - f_0\|_2 \leq \frac{\gamma}{2} \|x_0 - x^*\|_2^2 . \tag{50}$$

Using the triangular inequality then shows that

$$\left\|J_0^T f_0\right\|_2 \leq (\sigma + \alpha^2) \|x_0 - x^*\|_2 + \frac{\alpha\gamma}{2} \|x_0 - x^*\|_2^2 . \tag{51}$$

Gathering the partial results (47) and (51), we obtain

$$\begin{aligned}
\|x_1 - x^*\|_2 &= \left\|x_0 - (J_0^T J_0)^{-1} J_0^T f_0 + (J_0^T J_0)^{-1} r_0 - x^*\right\|_2 \\
&\leq \left\|x_0 - (J_0^T J_0)^{-1} J_0^T f_0 - x^*\right\|_2 + \|r_0\|_2 \left\|(J_0^T J_0)^{-1}\right\|_2 \\
&\leq \left\|x_0 - (J_0^T J_0)^{-1} J_0^T f_0 - x^*\right\|_2 + \beta_0 \left\|(J_0^T J_0)^{-1}\right\|_2 \left\|J_0^T f_0\right\|_2 \\
&\leq \frac{c}{\lambda}(\sigma + \beta_0(\sigma + \alpha^2)) \|x_0 - x^*\|_2 + C \|x_0 - x^*\|_2^2 , \qquad (52)
\end{aligned}$$

where $C = c\alpha\gamma(1+\hat{\beta})/(2\lambda)$, which proves (46) in the case $k = 0$. Since $\|x_0 - x^*\|_2 < \varepsilon$ is assumed initially, it follows from (45) and (48) that

$$\|x_1 - x^*\|_2 \leq \left(\frac{c}{\lambda}(\sigma + \hat{\beta}(\sigma + \alpha^2)) + C\varepsilon\right) \|x_0 - x^*\|_2 \leq K \|x_0 - x^*\|_2 < \|x_0 - x^*\|_2 , \tag{53}$$

where $K = (\lambda + c(\sigma + \hat{\beta}(\sigma + \alpha^2)))/(2\lambda) < 1$. The convergence is then established by repeating the argument for $k = 1, 2, \ldots$. ∎

The theorem shows that to ensure the convergence of the TGN method, the relative residuals in the solution of the inner linear least square problem must be bounded in terms of the parameters σ, λ, and α. The theorem also establishes the rates of convergence of the method in various cases. These cases are discussed in section 5.5.

We remark that the convergence of the TGN method can be established under weaker conditions than we give here, as proved in [10]. Only linear rates of convergence can be derived under the weaker conditions, however, whereas quadratic rates of convergence can be shown in certain cases under the assumptions made here (see section 5.5).

5.3. Convergence of the PGN method (II). In the next theorem we consider the PGN iteration where an approximate Jacobian $\tilde{J}$ is used instead of J.

THEOREM 12. *Let the conditions of Theorem* 9 *hold, and let* $\tilde{J}(x)$ *be an approximation to* $J(x)$. *Let* c *be such that* $1 < c < \lambda/\sigma$. *Assume that*

$$0 \le \hat{\eta} < \frac{\lambda - c\sigma}{c(\sigma + \alpha^2)}. \tag{54}$$

Then there exists $\varepsilon > 0$ *such that, if* $\|x_0 - x^*\|_2 < \varepsilon$ *and if*

$$\left\| J(x_k)^T J(x_k) \left(J^+(x_k) - \tilde{J}^+(x_k) \right) f(x_k) \right\|_2 / \left\| J(x_k)^T f(x_k) \right\|_2 \le \eta_k \le \hat{\eta}, \quad k = 0, 1, \ldots, \tag{55}$$

the sequence of PGN iterates $\{x_k\}$ *converge to* x^*. *Additionally, the following inequality holds:*

$$\|x_{k+1} - x^*\|_2 \le \frac{c}{\lambda}(\sigma + \eta_k(\sigma + \alpha^2)) \|x_k - x^*\|_2 + C \|x_k - x^*\|_2^2, \tag{56}$$

where $C = c\alpha\gamma(1 + \hat{\eta})/(2\lambda)$.

Proof of Theorem 12. The PGN iteration takes the form $x_{k+1} = x_k + s_k$, where $s_k = -\tilde{J}^+(x_k) f(x_k)$. Therefore, using the notation of Theorem 11, we may consider the PGN method as a TGN method with the residual defined by

$$r_k = J(x_k)^T J(x_k) s_k + J(x_k)^T f(x_k) = J(x_k)^T J(x_k)(J^+(x_k) - \tilde{J}^+(x_k)) f(x_k).$$

The conclusion then follows directly from Theorem 11. ∎

We remark that Theorem 12 establishes the convergence of the PGN method to the fixed point x^* of the *exact* GN method. At the fixed point, the perturbed Jacobian $\tilde{J}$ must, therefore, be such that $\tilde{J}(x^*)^T f(x^*) = 0$ in order to be able to satisfy the conditions of the theorem; that is, at the fixed point x^* the null space of $\tilde{J}(x^*)^T$ must contain $f(x^*)$. In contrast the convergence results of Theorem 6 require only that a point $\tilde{x}^*$ exists such that $\tilde{J}(\tilde{x}^*)^T f(\tilde{x}^*) = 0$ and $\tilde{J}(\tilde{x}^*)$ is full rank.

5.4. Convergence of the TPGN method (II). In the following theorem we consider the TPGN iteration where an approximate Jacobian $\tilde{J}$ is used and the inner linear least squares problem (15) is not solved exactly on each outer step. The residuals in the inner normal equations at each outer iteration are assumed to be bounded such that

$$\|r_k\|_2 \le \beta_k \left\| \tilde{J}(x_k)^T f(x_k) \right\|_2, \tag{57}$$

where $\{\beta_k\}$ is a nonnegative forcing sequence. Sufficient conditions for the TPGN method to converge are then given as follows.

THEOREM 13. *Let the conditions of Theorem* 9 *hold, and let* $\tilde{J}(x)$ *be an approximation to* $J(x)$. *Let* c *be such that* $1 < c < \lambda/\sigma$. *Assume that* $\eta_k \le \hat{\eta} < (\lambda - c\sigma)/(c(\sigma + \alpha^2))$, *and select* $\beta_k, k = 0, 1, \ldots,$ *such that*

$$\begin{aligned} 0 \;\le\; \beta_k \le \Big(& \eta_k \left\| J(x_k)^T f(x_k) \right\|_2 - \left\| J(x_k)^T J(x_k)(J^+(x_k) - \tilde{J}^+(x_k)) f(x_k) \right\|_2 \Big) \\ & \cdot \Big(\left\| J(x_k)^T J(x_k)(\tilde{J}(x_k)^T \tilde{J}(x_k))^{-1} \right\|_2 \left\| \tilde{J}(x_k)^T f(x_k) \right\|_2 \Big)^{-1} \end{aligned} \tag{58}$$

for $k = 0, 1, \ldots$. *Then there exists* $\varepsilon > 0$ *such that, if* $\|x_0 - x^*\|_2 < \varepsilon$, *the sequence of PGN iterates* $\{x_k\}$ *satisfying* (57) *converges to* x^*. *Additionally, the following inequality holds:*

$$\|x_{k+1} - x^*\|_2 \leq \frac{c}{\lambda}(\sigma + \eta_k(\sigma + \alpha^2)) \|x_k - x^*\|_2 + C \|x_k - x^*\|_2^2, \tag{59}$$

where $C = c\alpha\gamma(1 + \hat{\eta})/(2\lambda)$.

Proof of Theorem 13. The TPGN iteration takes the form $x_{k+1} = x_k + s_k$, where $s_k = -\tilde{J}^+(x_k)f(x_k) + (\tilde{J}(x_k)^T\tilde{J}(x_k))^{-1}r_k$. Therefore, using the notation of Theorem 11, we may consider the TPGN method as a TGN method with the residual defined as

$$\tilde{r}_k = J(x_k)^T J(x_k)(J^+(x_k) - \tilde{J}^+(x_k))f(x_k) + J(x_k)^T J(x_k)(\tilde{J}(x_k)^T\tilde{J}(x_k))^{-1}r_k. \tag{60}$$

Then, provided the condition (57) holds, we have

$$\begin{aligned} \|\tilde{r}_k\|_2 &\leq \left\|J(x_k)^T J(x_k)(J^+(x_k) - \tilde{J}^+(x_k))f(x_k)\right\|_2 \\ &\quad + \left\|J(x_k)^T J(x_k)(\tilde{J}(x_k)^T\tilde{J}(x_k))^{-1}\right\|_2 \beta_k \left\|\tilde{J}(x_k)^T f(x_k)\right\|_2 \\ &\leq \eta_k \left\|J(x_k)^T f(x_k)\right\|_2. \end{aligned} \tag{61}$$

The conclusion then follows from Theorem 11. ∎

We remark that to ensure $\beta_k \geq 0$ we require that the relation given by equation (55) holds. This is simply the condition of Theorem 12 that guarantees the convergence of the PGN method in the case where the inner loop is solved exactly without truncation.

Theorem 13 gives conditions for the TPGN method to converge to the fixed point x^* of the *exact* GN method and is therefore more restrictive than the theorem developed in section 4. Here the allowable form of the perturbed Jacobian is constrained to satisfy $\tilde{J}(x^*)^T f(x^*) = J(x^*)^T f(x^*) = 0$ in order that the conditions of the theorem may be met. The theorem does, however, establish that the method converges with rates of convergence higher than linear in certain cases. These cases are discussed in the next section.

5.5. Rates of convergence of the approximate GN methods. From Theorems 11, 12, and 13, the expected convergence rates of the approximate GN methods may be established for various cases. The convergence rates are shown in (46), (56), and (59) for the TGN, the PGN, and the TPGN methods, respectively. These rates are dependent on the parameters σ, λ, and α, defined as in Theorem 9, and can be contrasted directly with the convergence rates of the exact GN method, given by (37). We observe the following.

1. *Linear convergence.* The theorems show that in general if the GN, TGN, PGN, and TPGN methods converge, then they converge linearly. In comparison with the exact GN algorithm, we see that the price paid for the inaccurate solution of the linear least squares problem in the inner step of the approximate methods is a degradation of the local linear rate of convergence.
2. *Superlinear convergence.* As previously noted, if $\sigma = 0$, which holds, for example, in the zero-residual case where $f(x^*) = 0$, the convergence of the exact GN method is quadratic [9, Corollary 10.2.2]. In this same case, if $\sigma = 0$ and if the forcing sequence $\{\beta_k\}$ satisfies $\lim_{k\to+\infty} \beta_k = 0$, then the convergence

rates of the approximate TGN and TPGN methods are superlinear. For the PGN method to converge superlinearly in this case, the sequence $\{\eta_k\}$ must satisfy $\lim_{k\to+\infty} \eta_k = 0$.

3. *Quadratic convergence.* From the proof of Theorem 11, we see that the convergence of the TGN method is quadratic if $\sigma = 0$ and if the normal equation residual is such that

$$\|r_k\|_2 \equiv \left\|J(x_k)^T J(x_k) s_k + J(x_k)^T f(x_k)\right\|_2 \leq C_1 \left\|J(x_k)^T f(x_k)\right\|_2^2$$

for some positive constant C_1. Similarly, in the case $\sigma = 0$, the PGN method converges quadratically if

$$\left\|J(x_k)^T J(x_k) \left(J^+(x_k) - \tilde{J}^+(x_k)\right) f(x_k)\right\|_2 \leq C_2 \left\|J(x_k)^T f(x_k)\right\|_2^2,$$

as does the TPGN method in this case if

$$\left\|(J(x_k)^T J(x_k))((J^+(x_k) - \tilde{J}^+(x_k))f(x_k) + (\tilde{J}(x_k)^T \tilde{J}(x_k))^{-1} r_k)\right\|_2 \leq C_3 \left\|J(x_k)^T f(x_k)\right\|_2^2$$

for positive constants C_2, C_3.

4. *Effect of nonlinearity.* Since $\lambda - c\sigma > 0$, we also see from the theorems that the allowable upper bound on the truncation decreases as σ increases. Since σ is a combined measure of the nonlinearity and the residual size in the problem, we see therefore that, in order to guarantee convergence of the approximate methods, the inner linearized equation must be solved more accurately when the problem is highly nonlinear or when there is a large residual at the optimal.

In section 6 we give numerical results demonstrating the convergence behavior of the approximate GN methods. The rates of convergence of the approximate methods are also illustrated for various cases.

5.6. Summary. In this section we have established theory ensuring local convergence of the GN, the TGN, the PGN, and the TPGN methods based on the theory of [9] for exact GN methods. The conditions for convergence derived in this section are less general than those of section 4 but enable the rates of convergence to be established. Numerical examples illustrating the results for the three approximate GN methods are shown in the next section, and in section 7 an application to a practical problem in data assimilation is presented.

6. Numerical example. We examine the theoretical results of sections 4 and 5 using a simple data assimilation problem from [13, Chapter 4]. The aim is to fit the solution of a discrete dynamical model to observations of the model state at two points in time. The system dynamics are described by the ordinary differential equation

$$\frac{dz}{dt} = z^2, \tag{62}$$

where $z = z(t)$. A second-order Runge–Kutta scheme is applied to the continuous equation to give the discrete nonlinear model

$$x^{n+1} = x^n + (x^n)^2 \Delta t + (x^n)^3 \Delta t^2 + \frac{1}{2}(x^n)^4 \Delta t^3, \tag{63}$$

where Δt denotes the model time step and $x^n \approx z(t_n)$ at time $t_n = n\Delta t$. The data assimilation problem is then defined to be

$$\min_{x^0} \phi(x) = \frac{1}{2}(x^0 - y^0)^2 + \frac{1}{2}(x^1 - y^1)^2 \tag{64}$$

subject to (63), where $y^0,\ y^1$ are values of observed data at times $t_0,\ t_1$, respectively. This is a NLSP of the form (1) with

$$f = \begin{pmatrix} x^0 - y^0 \\ x^1 - y^1 \end{pmatrix}. \tag{65}$$

The Jacobian of f is given by

$$J(x^0) = \begin{pmatrix} 1 \\ 1 + 2x^0\Delta t + 3(x^0)^2\Delta t^2 + 2(x^0)^3\Delta t^3 \end{pmatrix}, \tag{66}$$

and the second-order terms of the Hessian are

$$Q(x^0) = \left(x^0 + (x^0)^2\Delta t + (x^0)^3\Delta t^2 + \tfrac{1}{2}(x^0)^4\Delta t^3 - y^1\right)(2\Delta t + 6x^0\Delta t^2 + 6(x^0)^2\Delta t^3). \tag{67}$$

We use this example to test the convergence theory for the approximate GN methods. In the experiments we set the true value of x^0 to -2.5 and begin the iteration with an initial estimate of -2.3 for x^0. The observations are generated by solving the discrete numerical model (63) with the true initial state. The time step is set to $\Delta t = 0.5$. The algorithms are considered to have converged when the difference between two successive iterates is less than 10^{-12}, and we restrict the maximum number of iterations to 1000. We first test the convergence of the TGN algorithm.

6.1. TGN method: Numerical results. The exact GN method is easy to apply to this simple example since we can solve the inner step directly at each iteration. In order to test the theory for the TGN algorithm, we apply an error to the exact GN step and solve instead the approximate equation

$$J(x_k^0)^T J(x_k^0) s_k = -J(x_k^0)^T f(x_k^0) + r_k, \tag{68}$$

where on each iteration we select the size of the residual r_k. We choose

$$r_k = \epsilon \left(\frac{\hat{\beta} - |Q(x_k^0)(J(x_k^0)^T J(x_k^0))^{-1}|}{1 + |Q(x_k^0)(J(x_k^0)^T J(x_k^0))^{-1}|} \right) |\nabla \phi(x_k^0)|, \tag{69}$$

with ϵ a specified parameter and $\hat{\beta} = 0.999$. From Theorem 5 we expect the algorithm to converge to the correct solution for values of ϵ less than 1. In Table 1 we show the results of the iterative process for various levels of truncation. The first and second columns of the table give the values of ϵ chosen for the truncation and the number of iterations taken to reach convergence. The third and fourth columns show the differences between the iterated and true solutions and the gradient of the objective function at the iterated solution, which should be 0 if the true minimum has been reached.

For $\epsilon = 0$ (the exact GN method) the exact solution is found in 5 iterations. As the value of ϵ is increased, the number of iterations to reach convergence also increases. At $\epsilon = 0.95$ the number of iterations needed to achieve convergence is 401, but even for this large truncation, the correct solution to the NLSP is attained, as seen from

TABLE 1
Perfect observations, exact Jacobian.

ϵ	Iterations	Error	Gradient
0.00	5	0.000000e+00	0.000000e+00
0.25	20	9.015011e-14	1.364325e-13
0.50	37	7.207568e-13	1.092931e-12
0.75	84	2.246647e-12	3.407219e-12
0.90	210	8.292034e-12	1.257587e-11
0.95	401	1.857048e-11	2.816403e-11
1.00	1000	3.143301e-04	4.765072e-04
1.05	431	2.652062e-02	3.880614e-02
1.10	231	5.357142e-02	7.568952e-02
1.15	163	8.101821e-02	1.106474e-01
1.20	130	1.093852e-01	1.444877e-01
1.25	112	1.394250e-01	1.781241e-01

TABLE 2
Imperfect observations, exact Jacobian.

ϵ	Iterations	Error	Gradient
0.00	10	4.440892e-15	7.778500e-15
0.25	17	9.503509e-14	1.806853e-13
0.50	32	6.181722e-13	1.176347e-12
0.75	66	1.671552e-12	3.180605e-12
0.90	128	4.250822e-12	8.088735e-12
0.95	181	6.231016e-12	1.185694e-11
1.00	359	1.052936e-11	2.003732e-11
1.05	157	6.324736e-02	1.093406e-01
1.10	116	8.697037e-02	1.452842e-01
1.15	93	1.103473e-01	1.783861e-01
1.20	79	1.336149e-01	2.092708e-01
1.25	69	1.570351e-01	2.384890e-01

the size of the error and gradient values. For a value of $\epsilon = 1.0$ the algorithm fails to converge within 1000 iterations. For values of ϵ greater than 1, the algorithm also converges, but the solution is not correct. With $\epsilon = 1.05$, for example, the stopping criterion is satisfied after 431 iterations, but the final gradient is now of the order 10^{-2}, indicating that a true minimum has not been found. Thus from these results it appears that the bound on the truncation proposed in Theorem 5 is precise. For truncations less than this bound we converge to the correct solution of the NLSP but not for truncations above this bound.

We note that, for this example, if the observations are perfect, then we have a zero-residual NLSP problem. In order to test the theory when this is not the case, we consider an example where the observational data contains errors, as generally occurs in practice. We add an error of 5% to observation y^0 and subtract an error of 5% from observation y^1. The true solution is then calculated by applying the full Newton method to the problem, giving the value $x^0 = -2.5938$ in 7 iterations. (The accuracy of this result is checked by ensuring that the gradient is 0 at the solution.) The convergence results for this test case are shown in Table 2, where the third column is now the difference between the iterated TGN solution and the solution calculated using the exact Newton method.

We see a similar pattern of behavior to that in the perfect observation (zero-residual) case. For all values of ϵ less than 1 the TGN algorithm converges to the same solution as the exact Newton method, but the number of iterations required for

convergence increases as ϵ increases. Where $\epsilon = 1$, the method now converges within the iteration limit to the optimal found by the Newton method. The procedure also converges for values of ϵ greater than 1, but the solution is no longer correct. In these cases the size of the gradient indicates that a minimum has not been found.

6.2. PGN method: Numerical results. In data assimilation, a perturbed Jacobian is often derived by replacing the linearized discrete dynamical model by a simplified discrete form of the linearized continuous system. In this test case we generate a perturbed Jacobian in the same way, following the example of [13]. The linearization of the continuous nonlinear equation (62) is written

$$\frac{d(\delta z)}{dt} = 2z(\delta z), \tag{70}$$

where $\delta z(t)$ is a perturbation around a state $z(t)$ that satisfies the nonlinear equation (62). Applying the second-order Runge–Kutta scheme to this equation gives

$$(\delta x)^{n+1} = \left(1 + 2x^n\Delta t + 3(x^n)^2\Delta t^2 + 3(x^n)^3\Delta t^3 + \tfrac{5}{2}(x^n)^4\Delta t^4 + (x^n)^5\Delta t^5\right)(\delta x)^n, \tag{71}$$

where $(\delta x)^n \approx \delta z(t_n)$ at time $t_n = n\Delta t$ and $x^n \approx z(t_n)$ satisfies the discrete nonlinear model (63). The perturbed Jacobian for this example thus becomes

$$\tilde{J}(x^0) = \begin{pmatrix} 1 \\ 1 + 2x^0\Delta t + 3(x^0)^2\Delta t^2 + 3(x^0)^3\Delta t^3 + \frac{5}{2}(x^0)^4\Delta t^4 + (x^0)^5\Delta t^5 \end{pmatrix}. \tag{72}$$

Using this perturbed Jacobian we apply the PGN algorithm to our example and test whether the sufficient condition (27) for convergence is satisfied on each iteration. For this example the second-order terms $\tilde{Q}$ are given by

$$\begin{aligned}\tilde{Q}(x^0) = &\left(x^0 + (x^0)^2\Delta t + (x^0)^3\Delta t^2 + \tfrac{1}{2}(x^0)^4\Delta t^3 - y^1\right) \\ &\cdot (2\Delta t + 6x^0\Delta t^2 + 9(x^0)^2\Delta t^3 + 10(x^0)^3\Delta t^4 + 5(x^0)^4\Delta t^5).\end{aligned} \tag{73}$$

In the case where we have perfect observations, condition (27) is satisfied on each iteration, and the PGN method converges to the true solution in 27 iterations. When error is added to the observations, as in the previous section, the PGN method converges in 21 iterations, and again, the condition for convergence is always satisfied. Now, however, the NLSP problem is no longer a zero-residual problem, and the fixed points of the GN and PGN methods are no longer the same. The converged solutions differ in the norm by 0.05389, which is within the minimum upper bound $\|J^+(\tilde{x}^*)f(\tilde{x}^*)\|_2 = 0.05425$ on the error (with $\nu = 0$) given by Theorem 7.

In order to examine a case in which the sufficient condition (27) is not satisfied on each iteration, we change the time step to $\Delta t = 0.6$, keeping all other parameters of the problem the same. With this time step the perturbed linear model has significantly different characteristics from the exact linear model (see [13]). The Jacobians J and $\tilde{J}$ are therefore also very different, as are the second derivative matrices Q and $\tilde{Q}$, which are dependent here upon the observations. For perfect observations (zero-residual problem), the PGN iterations converge to the same solution as the exact GN and Newton procedures in 36 iterations (compared with 10 and 7 iterations for the GN and Newton methods, respectively). The condition for convergence of the PGN method is satisfied on each iteration, with the maximum value of the left-hand side of (27) reaching 0.709. In the case where there are errors in the observed values,

TABLE 3
Imperfect observations, inexact Jacobian.

ϵ	Iterations	Error	Residual
0.00	21	8.215650e-14	6.693951e-14
0.25	33	4.911627e-13	4.007662e-13
0.50	56	1.217249e-12	9.930633e-13
0.75	121	3.732126e-12	3.044658e-12
0.90	306	1.105871e-11	9.021988e-12
0.95	596	2.444178e-11	1.993989e-11
1.00	1000	1.260007e-01	9.382085e-02
1.05	90	1.714365e+00	1.765471e+00
1.10	53	1.842029e+00	1.934063e+00
1.15	36	1.940084e+00	2.069636e+00
1.20	25	2.019233e+00	2.184031e+00
1.25	23	2.085381e+00	2.283791e+00

however, the perturbed problem is no longer close to the NLSP, and the condition for convergence of the PGN method fails on every second iteration. As anticipated in this case, the PGN method fails to converge, even after 1000 iterations. By comparison, the exact GN algorithm, using the true Jacobian, converges in 8 iterations to the same solution of the NLSP as that obtained by Newton's method. This example illustrates the effects of approximations on the convergence properties of the GN method and demonstrates the limits on the convergence of the PGN method as established by the theory of sections 4 and 5.

6.3. TPGN method: Numerical results. Finally in this section we consider the case in which the PGN method is also truncated. Following the same method as in the previous two sections, we solve on each iteration the approximate equation

$$\tilde{J}(x_k^0)^T \tilde{J}(x_k^0) s_k = -\tilde{J}(x_k^0)^T f(x_k^0) + r_k, \tag{74}$$

where we select the residual r_k. We choose

$$r_k = \epsilon \left(\frac{\hat{\beta} - |1 - (\tilde{J}(x_k)^T J(x_k) + \tilde{Q}(x_k))(\tilde{J}(x_k)^T \tilde{J}(x_k))^{-1}|}{|(\tilde{J}(x_k)^T J(x_k) + \tilde{Q}(x_k))(\tilde{J}(x_k)^T \tilde{J}(x_k))^{-1}|} \right) |\nabla \phi(x_k^0)|, \tag{75}$$

where ϵ is a specified parameter and $\hat{\beta} = 0.999$. The other data are as before, with errors added to the observations. From Theorem 8 we expect the method to converge for values of $\epsilon < 1$. The true solution is calculated by applying the exact Newton method to the perturbed problem, giving a result in 5 iterations of $x^0 = -2.6477$. In Table 3 we present the convergence results for the TPGN method using various levels of truncation. The third column now shows the difference between the TPGN solution and the exact Newton method applied to the perturbed problem, and the fourth column gives the residual in the perturbed equation (14). We find that, as expected from the theory, the TPGN algorithm converges to the correct solution for values of $\epsilon < 1$. For values of $\epsilon > 1$ the algorithm converges to an incorrect solution. Thus it appears that the bound derived in Theorem 8 is robust.

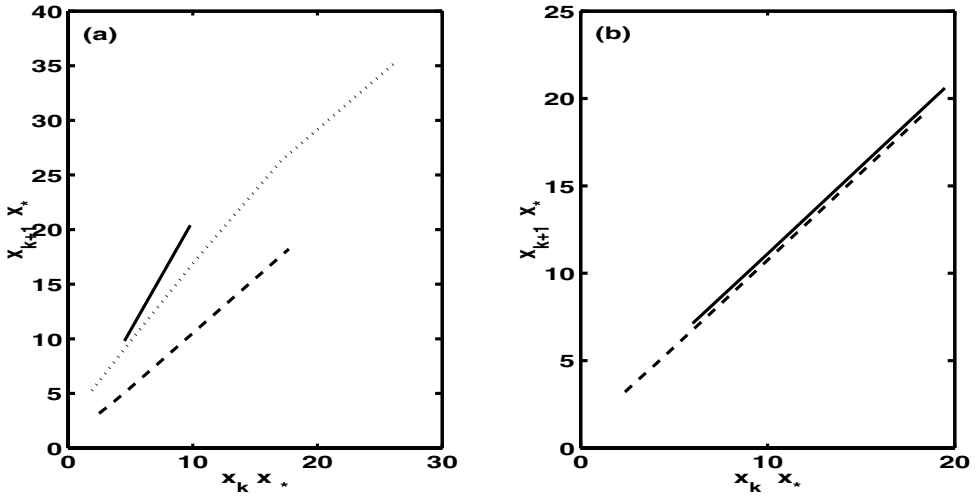


FIG. 1. *Convergence rates for the cases of* (a) *an exact Jacobian and* (b) *a perturbed Jacobian for the zero-residual case. The solid line is for no truncation; the dashed line is for constant truncation; and the dotted line in plot* (a) *is for variable truncation.*

6.4. Rates of convergence. Finally we test numerically the convergence rates derived in section 5. We examine the zero-residual problem for the numerical example with perfect observations. The convergence is measured in terms of the norms of the errors on each iteration. If the convergence is of order p, then

$$\| x_{k+1} - x^* \|_2 = K \| x_k - x^* \|_2^p \quad \text{for some constant } K, \tag{76}$$

where x^* is the exact fixed point. The plot of $|\log(\| x_{k+1} - x^* \|_2)|$ against $|\log(\| x_k - x^* \|_2)|$ will therefore have slope p.

In Figure 1(a) we plot this slope for the case where the exact Jacobian is used. Three curves are shown corresponding to the exact GN method, the TGN method with constant truncation, and the TGN method with $\beta_k \to 0$. From the theory of section 5 we expect the rates of the convergence for these cases to be quadratic, linear, and superlinear. We find that this is the case. For the exact GN method the slope of the line in the figure is 1.97, and for the TGN method it is 0.98. For the case in which the truncation tends to 0, the slope is 0.96 on the upper part of the line, which corresponds to the initial iterations, but steepens to 1.5 on the lower part of the line, demonstrating the superlinear nature of the convergence.

In Figure 1(b) the slope is plotted for the case where the perturbed Jacobian is used. We show the convergence for the PGN method with no truncation and the TPGN method with constant truncation. From the previous theory we expect both of these to have linear convergence. The numerical results show that this is the case, with both lines in the figure having a slope of one.

We conclude from these studies that the theoretical results of sections 4 and 5 predict the convergence behavior of the approximate GN methods reliably and robustly.

7. Application to data assimilation. We now consider the application of the approximate GN theory to a more realistic problem in atmosphere and ocean data assimilation. In the technique of 4D-Var the aim is to find an initial state x_0 at time

t_0 such that the distance between the trajectory of the numerical model of the system initiated from state x_0 and a set of observations of the system at times $t_i, i = 0, \ldots, N$, is minimized. In practice other constraints may also be added to the system so as to ensure, for example, that the state x_0 lies close to an a priori estimate. These constraints are not needed, however, in order to illustrate the theory developed here. We therefore express the 4D-Var problem in the form

$$\min_{x_0} \mathcal{J}[x_0] = \frac{1}{2} \sum_{i=0}^{N} (H_i[x_i] - y_i)^T R_i^{-1} (H_i[x_i] - y_i) \tag{77}$$

subject to $x_i = S(t_i, t_0, x_0)$, where S is the solution operator of the discrete nonlinear model, H_i is an operator that maps the model state into the p_i-dimensional vector of observations y_i at time t_i, and R_i is a weighting matrix given by the error covariance matrix of the observations at time t_i. This is an NLSP of the form (1) with

$$f(x_0) = - \begin{pmatrix} R_0^{1/2}(H_0[x_0] - y_0) \\ \vdots \\ R_N^{1/2}(H_N[x_N] - y_N) \end{pmatrix}, \tag{78}$$

where we note that the definition of $f(x_0)$ implicitly involves the nonlinear model operators $S(t_i, t_0, x_0)$ for calculating the states of the system x_i at the different observation times. In atmospheric data assimilation the size of the state vector x is very large, of order 10^7–10^8, and so efficient methods for minimizing (77) must be found. A common implementation of 4D-Var is the incremental formulation of [6], which minimizes a series of linear approximations to (77). Recently we have shown this to be equivalent to applying a GN method to the nonlinear problem (77) [14, 15].

7.1. Model problem. To illustrate the theory we examine a 4D-Var assimilation problem for a simplified model of fluid flow over an obstacle. The system is described by the one-dimensional nonlinear shallow water equations in the absence of rotation given by

$$\frac{\mathrm{D}u}{\mathrm{Dt}} + \frac{\partial \phi}{\partial \xi} = -g \frac{\partial \bar{h}}{\partial \xi}, \tag{79}$$

$$\frac{\mathrm{D}(\ln \phi)}{\mathrm{Dt}} + \frac{\partial u}{\partial \xi} = 0, \tag{80}$$

with

$$\frac{\mathrm{D}}{\mathrm{Dt}} = \frac{\partial}{\partial t} + u \frac{\partial}{\partial \xi}. \tag{81}$$

In these equations u is the wind velocity; $\phi = gh$ is the geopotential, where g is the gravitational constant; $h > 0$ is the height of the fluid above the topography; and $\bar{h}$ is the height of the underlying topography. We define the problem on a spatial domain $\xi \in [0, L]$ with periodic boundary conditions. The system is discretized using a two-time-level semi-implicit semi-Lagrangian integration scheme, as described in [16]. This is a similar numerical scheme to that commonly used in operational weather forecasting models (see, for example, [5, 7]).

We note that to apply the GN method to (77) we require the Jacobian of the function f, which includes the linearization of the discrete nonlinear model operator

S. This linear operator is known as the tangent linear model and is usually found by a direct linearization of the discrete nonlinear model, using the procedure of automatic differentiation [1]. An alternative method of generating the linear model is first to linearize the continuous nonlinear equations (79) and (80) and then to discretize the resulting linear equations. This is the method that the UK Met Office has used to develop the linear model for their operational assimilation scheme [18]. In general this method will usually give a discrete model that is not the exact linearization of the discrete nonlinear model. Thus when this model is used within the GN iteration, we have only an approximate Jacobian $\tilde{J}$. The theory for the PGN method then applies.

To demonstrate this technique with the shallow water model, we let $\delta u(\xi, t)$, $\delta\phi(\xi, t)$ denote perturbations around states $\bar{u}(\xi, t)$, $\bar{\phi}(\xi, t)$ that satisfy the nonlinear equations. Then substituting into (79) and (80) gives the linearized equations

$$\frac{\mathrm{D}(\delta u)}{\mathrm{Dt}} + (\delta u)\frac{\partial \bar{u}}{\partial \xi} + \frac{\partial(\delta\phi)}{\partial \xi} = 0, \tag{82}$$

$$\frac{\mathrm{D}}{\mathrm{Dt}}\Big(\frac{\delta\phi}{\bar{\phi}}\Big) + (\delta u)\frac{\partial(\ln\bar{\phi})}{\partial \xi} + \frac{\partial(\delta u)}{\partial \xi} = 0, \tag{83}$$

where

$$\frac{\mathrm{D}}{\mathrm{Dt}} = \frac{\partial}{\partial t} + \bar{u}\frac{\partial}{\partial \xi} \tag{84}$$

is the material derivative defined with the linearization state wind $\bar{u}$. These equations are discretized using a semi-implicit semi-Lagrangian scheme, similar to that used in the full nonlinear model. Full details of the resulting numerical scheme are given in [16], where it is also shown that this scheme is different from that of the tangent linear model.

To solve the inner step of the GN method, a conjugate gradient iterative minimization is applied to the exact or perturbed linear least squares problem (6) or (15). The gradient directions are determined using the adjoints of the discrete linear model equations. We note that, in the case where the inexact linear model is used, this procedure generates the perturbed gradient $\tilde{J}(x)^T f(x)$, as required by the iteration method, rather than the exact gradient $J(x)^T f(x)$.

The inner iteration is stopped when the relative residual, defined as in (11) or (16), falls below a specified tolerance. If the minimization is stopped before full convergence, then the theory derived in sections 4 and 5 for the convergence of the TGN or TPGN method applies. For a complex model such as this, it is not possible to calculate the tolerances on the residuals required by the theory. The convergence results do, however, provide a theoretical basis for a practical inner loop stopping criterion that leads to smoother convergence of the outer loops and more efficient algorithms for solving the data assimilation problem, as described in [17]. Suitable tolerances are found experimentally, and the inner loop is stopped when the relative change in the gradient of the linear least squares problem, which is equal to the relative residual, is below the selected tolerance.

We perform idealized assimilation experiments in which the observations are generated by the nonlinear model, starting from a known initial state that we define to give the truth. These observations are then used in the assimilation, which is started from an incorrect prior estimate of the state. Further details of the implementation of the assimilation scheme can be found in [14]. Assimilation experiments are performed

with the exact linear model (and hence exact Jacobian) and with the inexact linear model (perturbed Jacobian). We consider the cases of both perfect observations, which imply a zero-residual problem, and observations with random Gaussian error added, which lead to a nonzero residual. We see that in practice the convergence behavior of the approximate GN methods is as predicted by the theory established in sections 4 and 5.

7.2. Numerical results. For the numerical experiments we consider a periodic domain of 200 grid points with a spacing of $\Delta\xi = 0.01$ m, so that $\xi \in [0\text{ m}, 2\text{ m}]$. In the center of the domain we define an idealized mountain by the formula

$$\bar{h}(\xi) = \bar{h}_c \left(1 - \frac{\xi^2}{a^2}\right) \quad \text{for} \quad 0 < |\xi| < a \tag{85}$$

and $\bar{h}(\xi) = 0$, otherwise, where we choose $\bar{h}_c = 0.05$ m and a is taken to be $40\Delta\xi = 0.4$ m. The gravitational constant g is set to be 10 ms^{-2}, and the model time step is 9.2×10^{-3} s.

Idealized observations of both u and ϕ are taken at every spatial grid point ξ_j, $j = 1, \ldots, 200$, over a time window of 50 steps. This gives a state vector of dimension $n = 400$ with $p_i = 400$ observations of the states at all grid points at times t_i, $i = 0, \ldots, 50$. The function f is thus of dimension $m = 20400$. The observations are assimilated using the GN or PGN method, starting from an incorrect prior estimate, which is generated by adding a phase shift of 0.5 m to the true initial state. The outer iteration is stopped when the norm of the gradient or the perturbed gradient, defined as in (5) or (12), respectively, is less than a given tolerance. The tolerance is set to 0.005 in the case of perfect observations and to 0.05 in the case where random errors have been added to the observations.

The inner minimization loop is stopped when the relative change in the gradient of the linear least squares problem falls below tolerances of 0.1 and 0.9. For a tolerance of 0.1 we find that the inner loop is fully converged and choosing a value several orders of magnitude lower than this does not change the convergence of the GN or PGN iterations. Thus we can consider this case as the method without truncation. The value of 0.9 on the other hand corresponds to a severely truncated method. In the zero-residual case for the exact Jacobian, we also perform a variable truncation experiment in which the truncation tolerance is initially set to 0.9 but is halved on each new GN iteration.

We consider first the case where we have exact observations of the true solution, so that the NLSP has a zero residual. In Figure 2 we plot the values of the nonlinear cost function (77) and the gradient, or perturbed gradient, for each iterative procedure. We observe that the fastest convergence is obtained when using the exact Jacobian with no truncation, as we expect from the theory. When the truncation level is increased, then the convergence rate slows down. (We remark that as the tolerance level is increased, however, less work is needed per iteration.) From Figure 2 we see also that if the tolerance tends to 0 as the iterations proceed, then the convergence rate increases. This behavior is expected from the theory, which predicts superlinear convergence in this case.

When the perturbed Jacobian is used, the convergence rate is found to be the same or slightly slower than when using the exact Jacobian, depending on whether truncation is applied or not. In general for the perturbed Jacobian with no truncation, this rate is faster than we would expect from the theory, which predicts only linear convergence. As we showed in section 5.5, however, the PGN method can converge

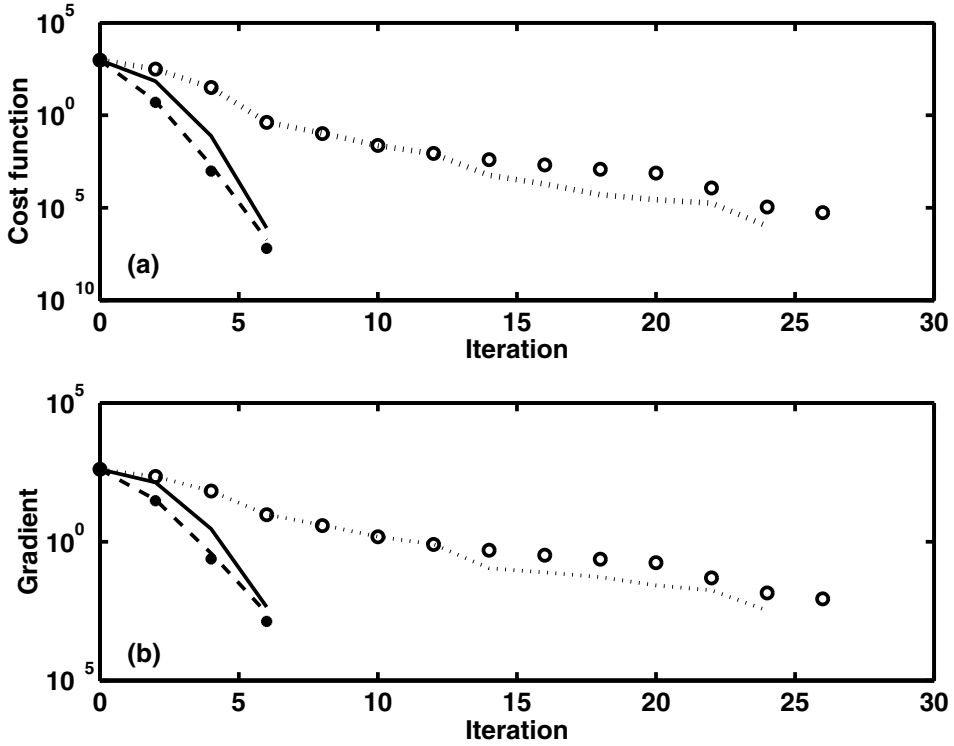


FIG. 2. *Convergence of the* (a) *cost function and* (b) *gradient, or perturbed gradient, for the perfect observation case. The different lines are for the exact Jacobian with no truncation (dashed line), with truncation (dotted line), and with variable truncation (solid line). The symbols indicate the convergence for the perturbed Jacobian with no truncation (solid dots) and with truncation (circles).*

quadratically for the zero-residual problem if the perturbed and exact Jacobians are close to each other. This can explain the fast convergence with the inexact Jacobian in this case.

For the experiments with imperfect observations, the observational errors for u and ϕ are taken from a Gaussian distribution, with standard deviations equal to 5% of the mean value of each field. The rates of convergence for these experiments are shown in Figure 3. In this case the solution does not fit the observations exactly, and we have a nonzero-residual problem. The fixed points of the exact GN and the PGN iterations are not the same now and differ in the norm by 0.01425. This difference is within the minimum upper bound $\|J^+(\tilde{x}^*)f(\tilde{x}^*)\|_2 = 0.01608$ on the error (with $\nu = 0$) predicted by Theorem 7.

For the nonzero-residual problem we expect linear convergence for all cases and so we do not show a variable truncation run. If we examine the convergence for the exact Jacobian, we again find that, where truncation is used on the inner minimization, the convergence rate slows down in comparison with the case where the inner problem is solved exactly. However, the difference is less marked than in the zero-residual case. This is as expected from Theorem 11, which predicts linear convergence for both of these cases, but with the rate constant increasing with the degree of truncation. When the inexact Jacobian is used we see that, as for the perfect observation case, the convergence is very close to that using the exact Jacobian.

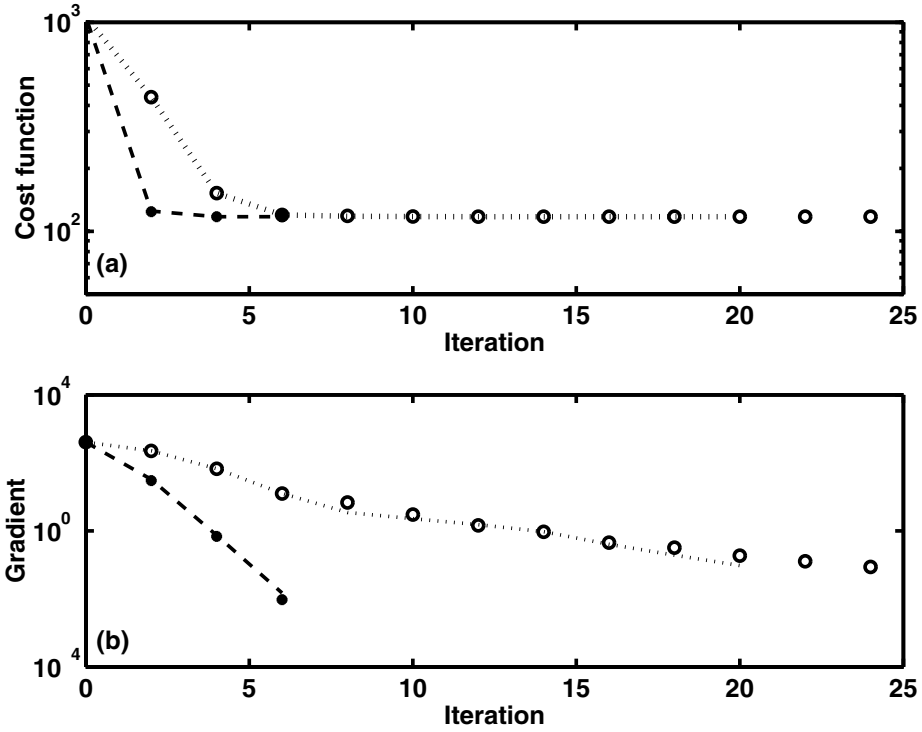


FIG. 3. *Convergence of the* (a) *cost function and* (b) *gradient, or perturbed gradient, for the imperfect observation case. The lines are for the exact Jacobian with no truncation (dashed line) and with truncation (dotted line). The symbols indicate the convergence for the perturbed Jacobian with no truncation (solid dots) and with truncation (circles).*

8. Conclusions. We have described here three approximate GN methods, the TGN, the PGN, and the TPGN methods, for solving the NLSP. We have derived conditions for the local linear convergence of these approximate methods by treating them as IN methods, following the theory of [8]. Additional, more restricted, convergence results have been derived for the approximate methods by extending the theory of [9] for the exact GN method. Through this approach, higher-order rates of convergence have been established for the approximate methods. We remark that many of these results could be generalized to hold under weaker assumptions than we have made here. The theory for IN methods could also be applied to give results on higher convergence rates for the approximate methods (see [4] and [8]).

In practice the approximate GN methods are used to treat very large data assimilation problems arising in atmosphere and ocean modeling and prediction. The convergence properties of these algorithms have not previously been investigated. Here we have established sufficient conditions for convergence to hold, allowing the approximate methods to be used operationally with confidence. By a simple numerical example we have shown that the bounds established by the theory are precise, in a certain sense, and that the approximate methods are convergent if the conditions of the theory hold. We have also demonstrated the application of the theory to a data assimilation problem for a typical model of a meteorological system.

Acknowledgment. We are grateful to Professor Gene Golub of Stanford University for support that enabled the completion of this work *in real time*.

REFERENCES

[1] M. C. Bartholomew-Biggs, S. Brown, B. Christianson, and L. Dixon, *Automatic differentiation of algorithms*, J. Comput. Appl. Math., 124 (2000), pp. 171–190.

[2] Åke Björck, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia, PA, 1996.

[3] E. Catinas, *Inexact perturbed Newton methods and applications to a class of Krylov solvers*, J. Optim. Theory Appl., 108 (2001), pp. 543–571.

[4] E. Catinas, *On the superlinear convergence of the successive approximations method*, J. Optim. Theory Appl., 113 (2002), pp. 473–485.

[5] J. Côté, S. Gravel, A. Méthot, A. Patoine, M. Roch, and A. Staniforth, *The operational CMC-MRB global environmental multiscale (GEM) model. Part* I: *Design considerations and formulation*, Monthly Weather Rev., 126 (1998), pp. 1373–1395.

[6] P. Courtier, J. N. Thepaut, and A. Hollingsworth, *A strategy for operational implementation of* 4*D-Var, using an incremental approach*, Quart. J. Roy. Meteor. Soc., 120 (1994), pp. 1367–1387.

[7] M. J. P. Cullen, T. Davies, M. H. Mawson, J. A. James, S. C. Coulter, and A. Malcolm, *An Overview of Numerical Methods for the Next Generation U.K. NWP and Climate Model*, Numer. Methods Atmos. Oceanic Model., The Andre Robert Memorial Volume, C. A. Lin, R. Laprise, and H. Ritchie, eds., Canadian Meteorological and Oceanographic Society, Ottawa, ON, Canada, 1997, pp. 425–444.

[8] R. S. Dembo, S. C. Eisenstat, and T. Steihaug, *Inexact Newton methods*, SIAM J. Numer. Anal., 19 (1982), pp. 400–408.

[9] J. E. Dennis and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice–Hall, Englewood Cliffs, NJ, 1983.

[10] J. E. Dennis and T. Steihaug, *On the successive projections approach to least squares problems*, SIAM J. Numer. Anal., 23 (1986), pp. 717–733.

[11] M. Ghil and P. Malanotte-Rizzoli, *Data assimilation in meteorology and oceanography*, Adv. Geophys., 33 (1991), pp. 141–266.

[12] S. Gratton, *Outils Théoriques d'Analyse du Calcul à Précision Finie*, Ph.D. thesis, TH/PA/98/30, Institut National Polytechnique de Toulouse, Toulouse, France, 1998.

[13] A. S. Lawless, *Development of linear models for data assimilation in numerical weather prediction*, Ph.D. thesis, The University of Reading, Reading, UK, 2001.

[14] A. S. Lawless, S. Gratton, and N. K. Nichols, *An investigation of incremental* 4*D-Var using non-tangent linear models*, Quart. J. Roy. Meteor. Soc., 131 (2005), pp. 459–476.

[15] A. S. Lawless, S. Gratton, and N. K. Nichols, *Approximate iterative methods for variational data assimilation*, Internat. J. Numer. Methods Fluids, 47 (2005), pp. 1129–1135.

[16] A. S. Lawless, N. K. Nichols, and S. P. Ballard, *A comparison of two methods for developing the linearization of a shallow-water model*, Quart. J. Roy. Meteor. Soc., 129 (2003), pp. 1237–1254.

[17] A. S. Lawless and N. K. Nichols, *Inner loop stopping criteria for incremental four-dimensional variational data assimilation*, Monthly Weather Rev., 134 (2006), pp. 3425–3435.

[18] A. C. Lorenc, S. P. Ballard, R. S. Bell, N. B. Ingleby, P. L. F. Andrews, D. M. Barker, J. R. Bray, A. M. Clayton, T. Dalby, D. Li, T. J. Payne, and F. W. Saunders, *The Met. Office global* 3*-dimensional variational data assimilation scheme*, Quart. J. Roy. Meteor. Soc., 126 (2000), pp. 2991–3012.

[19] N. K. Nichols, *Data assimilation: Aims and basic concepts*, in Data Assimilation for the Earth System, R. Swinbank, V. Shutyaev, and W. A. Lahoz, eds., Kluwer Academic Publishers, Norwell, MA, 2003, pp. 9–20.

[20] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[21] V. Pereyra, *Iterative methods for solving nonlinear least squares problems*, SIAM J. Numer. Anal., 4 (1967), pp. 27–36.

[22] J. S. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, Springer, New York, 1980.

[23] P-Å. Wedin, *On the Gauss–Newton Method for the Nonlinear Least-Squares Problems*, Working paper 24, Institute for Applied Mathematics, Stockolm, Sweden, 1974.

SIAM J. OPTIM.
Vol. 18, No. 1, pp. 133–137

REPRESENTATION OF SETS OF LATTICE POINTS*

RAYMOND HEMMECKE† AND ROBERT WEISMANTEL†

Abstract. Original algorithmic approaches in integer programming rely on the availability of different representations for all the lattice points in the feasible region. We present three results that are applicable to fairly general sets of lattice points and characterize a nonnegative integer linear representation.

Key words. integral generating set, Hilbert basis, representation, lattice, optimality certificate

AMS subject classification. 90C10

DOI. 10.1137/040616474

1. Introduction.

A fundamental theorem for the theory of linear integer optimization states that for every rational polyhedral cone C there exists a finite subset H of the set of integer points in C such that every integer point in C can be represented as a nonnegative linear integer combination of elements in H. This result follows from the work of Gordan [3]; see also [7, 8]. It has applications in many scenarios for linear integer programming. In particular, it is important for

- proving finiteness results in cutting plane theory [7],
- showing the existence of totally dual integral systems of linear diophantine systems [2],
- deriving optimality conditions for linear integer optimization problems [4],
- designing integer simplex type methods based on reformulation techniques [5].

In this paper it is our goal to extend this result to general sets of lattice points. Here, and throughout the paper, $\mathbb{Z}_+$ denotes the nonnegative integer numbers. For a set $P \subseteq \mathbb{R}^n$ we denote by

$$\text{recCone}(P) := \{\delta \in \mathbb{R}^n : \exists x \in P \text{ such that } x + \lambda\delta \in P \text{ for all } \lambda \geq 0\}$$

the recession cone of P. Moreover, for $S \subseteq \mathbb{Z}^n$, we call $T \subseteq S$ an *integral generating set* of S if for every $s \in S$ there exists a finite (integer) linear combination $s = \sum \alpha_i t_i$ with $t_i \in T$ and $\alpha_i \in \mathbb{Z}_+$.

It follows immediately that not every set S has a finite integral generating set. The following general result establishes necessary and sufficient conditions for a given rational polyhedral cone $C = \text{cone}(V)$ such that a finite nonnegative integer representation of S with elements in $S \cup C$ is possible.

The result can be used to determine a minimal number of vectors V that have to be added to a given set S to guarantee the existence of an integral generating set of S.

*Received by the editors October 6, 2004; accepted for publication (in revised form) July 18, 2006; published electronically February 2, 2007. This work was supported by the European TMR network ADONET 504438.

http://www.siam.org/journals/siopt/18-1/61647.html

†Department of Mathematics, Institute for Mathematical Optimization (IMO), Otto-von-Guericke-Universität Magdeburg, Universitätsplatz 2, 39106 Magdeburg, Germany (hemmecke@imo.math.uni-magdeburg.de, weismantel@imo.math.uni-magdeburg.de).

THEOREM 1.1.

(a) *Let $S \subseteq \mathbb{Z}^n$, and let $C = \operatorname{cone}(v_1, \ldots, v_m) \subseteq \mathbb{R}^n$ be a rational polyhedral cone. Then there exists a polytope $Q \subseteq \mathbb{R}^n$ with $S \subseteq Q + C$ if and only if there exists a finite set $H \subseteq S$ such that every $s \in S$ can be written as*

$$s = h + \sum_{j=1}^{m} \alpha_j v_j$$

for some $h \in H$ and with $\alpha_j \in \mathbb{Z}_+$, $j = 1 \ldots, m$.

(b) *Every nonempty set $S \subseteq \mathbb{Z}^n$ possesses a finite integral generating set if and only if the following two conditions hold:*
   - *The recession cone* $\operatorname{recCone}(\operatorname{conv}(S))$ *of* $\operatorname{conv}(S)$ *is a rational polyhedral cone.*
   - $\operatorname{recCone}(\operatorname{conv}(S)) \setminus \operatorname{conv}(S)$ *contains only finitely many lattice points.*

As a consequence of Theorem 1.1, we recover a few nice and well-known facts.

COROLLARY 1.2 (Giles and Pulleyblank [2]). *For every nonempty rational polyhedron $P \subseteq \mathbb{R}^n$, there exist a rational polytope $Q \subseteq \mathbb{R}^n$ and a rational polyhedral cone $C \subseteq \mathbb{R}^n$ such that $P \cap \mathbb{Z}^n = (Q \cap \mathbb{Z}^n) + (C \cap \mathbb{Z}^n)$.*

Proof. By Weyl's theorem, every rational polyhedron P can be written as the Minkowski sum of a polytope Q and the recession cone C of P. Thus, $S = P \cap \mathbb{Z}^n \subseteq Q + C$. From the first part of Theorem 1.1 we now get a desired representation $S = P \cap \mathbb{Z}^n = (\operatorname{conv}(H) \cap \mathbb{Z}^n) + (C \cap \mathbb{Z}^n)$. ∎

COROLLARY 1.3 (Jeroslow [6]). *A monoid $M \subseteq \mathbb{Z}^n$ has a finite (integral) generating set if and only if $C = \operatorname{cone}(M)$ is a rational polyhedral cone.*

Proof. If $C = \operatorname{cone}(M)$ is a rational polyhedral cone, then there exist finitely many elements $v_1, \ldots, v_m \in M$ that generate C. Therefore, we have $M \subseteq \{0\} + \operatorname{cone}(v_1, \ldots, v_m)$ and consequently, by the first part of Theorem 1.1, there exists a finite generating set $H \cup \{v_1, \ldots, v_m\}$ for M. The reverse direction is trivial, since any basis of M forms a generating set for $C = \operatorname{cone}(M)$. If the basis is finite, C is a rational polyhedral cone. ∎

As a generalization of Jeroslow's theorem one obtains the following statement, in which we do not assume a structure on the set S of lattice points.

COROLLARY 1.4. *Let $S \subseteq \mathbb{Z}^n$ be any set of lattice points in $\mathbb{Z}^n$. Then the following statements are equivalent:*

(1) *S has a finite integral generating set.*
(2) *$\operatorname{cone}(S)$ has a finite integral generating set.*
(3) *$\operatorname{conv}(S)$ has a finite integral generating set.*

Proof. (1) ⇔ (2): It is easy to see that S has a finite integral generating set if and only if the monoid generated by S has a finite integral generating set. The result follows now from Corollary 1.3 and by the fact that a polyhedral cone C possesses a finite integral generating set (Hilbert basis) if and only if C is rational [2].

(1) ⇒ (3): Let $V \subseteq S$ be a finite integral generating set for S. Then $\operatorname{cone}(S) = \operatorname{cone}(V)$ is a rational polyhedral cone and $\operatorname{conv}(S) \subseteq \{0\} + \operatorname{cone}(S) = \{0\} + \operatorname{cone}(V)$. Now, part (a) of Theorem 1.1 implies that there is some finite set $H \subseteq \operatorname{conv}(S) \cap \mathbb{Z}^n$ such that $H \cup V$ forms a finite integral generating set of $\operatorname{conv}(S)$.

(3) ⇒ (2): Let $V \subseteq \operatorname{conv}(S)$ be a finite integral generating set of $\operatorname{conv}(S)$. Then we conclude that $\operatorname{cone}(S) = \operatorname{cone}(\operatorname{conv}(S)) = \operatorname{cone}(V)$ is a rational polyhedral cone and thus possesses a finite integral generating set. ∎

Part (b) of Theorem 1.1 is an extension of the following result by using the equivalence (1) ⇔ (3) of Corollary 1.4.

COROLLARY 1.5 (Bertsimas and Weismantel [1]). *For $A \in \mathbb{Z}^{d\times n}$ and $b \in \mathbb{Z}^d$, define the sets $P = \{x \in \mathbb{R}^n_+ : Ax \le b\}$, $S = P \cap \mathbb{Z}^n$, and $C = \{x \in \mathbb{R}^n_+ : Ax \le 0\}$. If S is not empty, there exists a finite integral generating set of S if and only if S contains all but finitely many integer points in $C \cap \mathbb{Z}^n$.*

Theorem 1.1 allows us to classify arbitrary sets $S \subseteq \mathbb{Z}^n$ according to the minimum number of vectors $v_1, \dots, v_r \in \mathbb{Z}^n$ *not* in S such that $\text{cone}(S \cup \{v_1, \dots, v_r\})$ is a rational polyhedral cone. We call this number the polyhedral defect of $\text{cone}(S)$ or simply of S. Note that the polyhedral defect is always between 0 and $n+1$, since $n+1$ is the conic dimension of $\mathbb{R}^n$. Sets S have a polyhedral defect of 0 if and only if $\text{cone}(S)$ is rational and polyhedral.

It is an interesting open question to characterize all those sets $S \subseteq \mathbb{Z}^n$ that have a given polyhedral defect. As a subproblem, we would find it interesting to devise an algorithm that determines the polyhedral defect for sets $S = P \cap \mathbb{Z}^n$, where P is a given rational polyhedron.

If S is the set of integer points in a semialgebraic set, then Theorem 1.1 can be applied only if we know a polytope Q and a rational polyhedral cone C with $S \subseteq Q + C$. It is, however, a nontrivial task to determine such a cone C generated only by a minimal number of generators not belonging to S. This raises the question of whether there exist other representations for $x \in S$ of the kind

$$x = s_x + v_x,$$

where $s_x \in S$ and $v_x \in V$, where V is not necessarily of the form

$$V = \left\{ v \in \mathbb{Z}^n : v = \sum_{i=1}^{m} \alpha_i v_i, \alpha_i \in \mathbb{Z}_+, i = i, \dots, m \right\}$$

as in Theorem 1.1.

For a quite general family of semialgebraic sets, such a representation is possible. Interestingly, when the semialgebraic set describes a rational polyhedral cone, our proof recovers finiteness of a Hilbert basis.

THEOREM 1.6. *Let $\mathcal{C} := \{x \in \mathbb{R}^n : \exists y \ge 0$ with $x = g(y)\}$ be a semialgebraic set, where $g : \mathbb{R}^d \to \mathbb{R}^n$ is a vector of polynomial functions that map integer vectors to integers. There exist two vectors of polynomial functions $g_l, g_u : \mathbb{R}^n \to \mathbb{R}^n$ with $\text{maxdeg}(g_l), \text{maxdeg}(g_u) < \text{maxdeg}(g)$ such that for every point $x \in S = \mathcal{C} \cap \mathbb{Z}^n$ there exist a $\lambda \in \mathbb{Z}^d_+$ and a point $v_x \in \mathbb{Z}^n$ satisfying $x = g(\lambda) + v_x$ and $g_l(\lambda) \le v_x \le g_u(\lambda)$.*

2. Proofs of main theorems.

Proof of Theorem 1.1. If there exists a finite set $H \subseteq S$ such that every $s \in S$ can be written as

$$s = h + \sum_{j=1}^{m} \alpha_j v_j$$

for some $h \in H$ and with $\alpha_j \in \mathbb{Z}_+$, $j = 1 \dots, m$, then we have $S \subseteq \text{conv}(H) + C$.

It remains to show that given a polytope $Q \subseteq \mathbb{R}^n$ with $S \subseteq Q + C$, there exists a representation as desired.

First, note that $S \subseteq Q + C$ implies

$$S \subseteq \bigcup_{q \in Q \cap \mathbb{Z}^n} (q + C).$$

Next, let us triangulate C into (finitely many!) rational simplicial cones $C_1, \ldots, C_l$. Note that we can and do choose such a triangulation for which the generators of the cones C_i are also among the generators $v_1, \ldots, v_m$ of C. Then

$$S \subseteq \bigcup_{q \in Q \cap \mathbb{Z}^n} \bigcup_{i=1}^{l} (q + C_i).$$

We now construct a desired representation for each of the sets $(q + C_i) \cap S$. Without loss of generality, we assume $C_i = \operatorname{cone}(v_1, \ldots, v_r)$ (otherwise relabel the v_i). Consider the parallelepiped

$$F_i = \left\{ \sum_{j=1}^{r} \alpha_j v_j : 0 \leq \alpha_1, \ldots, \alpha_r < 1 \right\}.$$

As F_i is bounded, F_i contains only finitely many lattice points $\{f_1, \ldots, f_t\}$ in $\mathbb{Z}^n$. Moreover, $(q + C_i) \cap \mathbb{Z}^n$ is the disjoint union of the following t sets $F_{i,1}, \ldots, F_{i,t}$ with

$$F_{i,j} = \left\{ q + f_{i,j} + \sum_{k=1}^{r} \alpha_k v_k : \alpha_1, \ldots, \alpha_r \in \mathbb{Z}_+ \right\}.$$

Thus, it suffices to construct a desired representation for each of the sets $F_{i,j} \cap S$. If $q + f_{i,j} \in S$, we have already found a desired representation of the points $F_{i,j} \cap S$. Thus assume on the contrary that $q + f_{i,j} \notin S$. We construct now a finite set $H_{i,j} \subseteq S$ such that every $s \in F_{i,j} \cap S$ can be written as

$$s = h + \sum_{j=1}^{r} \alpha_j v_j$$

for some $h \in H_{i,j}$ and with $\alpha_j \in \mathbb{Z}_+$, $j = 1 \ldots, m$, as desired.

As C_i is a simplicial cone, each point in $F_{i,j}$ has a unique representation as $q + f_{i,j} + \sum_{k=1}^{r} \alpha_k v_k$, implying that there is a one-to-one correspondences $\phi_{i,j}$ between $F_{i,j}$ and $\mathbb{Z}_+^r$ given by

$$\phi_{i,j}\left(q + f_{i,j} + \sum_{k=1}^{r} \alpha_k v_k \right) = (\alpha_1, \ldots, \alpha_r).$$

Consider the set $\phi_{i,j}(F_{i,j} \cap S) \subseteq \mathbb{Z}_+^r$. By the Gordan–Dickson lemma [3], there are only finitely many points $\{g_1, \ldots, g_p\}$ that are minimal with respect to the partial ordering $\leq$ defined on $\mathbb{Z}_+^r$. Let $H_{i,j} = \{\phi_{i,j}^{(-1)}(g_1), \ldots, \phi_{i,j}^{(-1)}(g_p)\} \subseteq S$. Thus, for every element $s \in F_{i,j} \cap S$ there exists some $g \in \{g_1, \ldots, g_p\}$ with $g \leq \phi_{i,j}(s)$, implying that $s = \phi_{i,j}^{(-1)}(g) + \sum_{k=1}^{r} \alpha_k v_k$ for $\phi_{i,j}^{(-1)}(g) \in H_{i,j}$ and $\alpha_k = (\phi_{i,j}(s) - g)^{(k)} \in \mathbb{Z}_+$, $k = 1, \ldots, r$. □

Proof of Theorem 1.6. Choose any $x \in S := \mathcal{C} \cap \mathbb{Z}^n$. Then $x = g(y)$ for some $y \in \mathbb{R}_+^n$. Now define $\lambda := \lfloor y \rfloor$ componentwise and let $v_x = x - g(\lambda)$ and $h = y - \lambda$. We will now construct functions $g_l : \mathbb{R}^n \to \mathbb{R}^n$ and $g_u : \mathbb{R}^n \to \mathbb{R}^n$, with the desired properties.

Let $D = \operatorname{maxdeg}(g)$. By multivariate Taylor expansion, we get for $j = 1, \ldots, n$

$$x^{(j)} = g^{(j)}(\lambda + h) = g^{(j)}(\lambda) + \sum_{i=1}^{D} \frac{1}{i!} \cdot \sum_{\substack{\alpha \in \mathbb{Z}_+^n : \\ \|\alpha\|_1 = i}} \frac{\mathrm{d}g^{(j)}(\lambda)}{\mathrm{d}x^{\alpha}} \cdot h^{\alpha}.$$

Therefore,

$$v_x^{(j)} = x^{(j)} - g^{(j)}(\lambda) = \sum_{i=1}^{D} \frac{1}{i!} \cdot \sum_{\substack{\alpha \in \mathbb{Z}_+^n : \\ \|\alpha\|_1 = i}} \frac{\mathrm{d}g^{(j)}(\lambda)}{\mathrm{d}x^\alpha} \cdot h^\alpha.$$

Note that $\operatorname{maxdeg}(\frac{\mathrm{d}g^{(j)}}{\mathrm{d}x^\alpha}) < \operatorname{maxdeg}(g)$ and that $0 \le h < 1$ by construction.

This sum is a polynomial in λ and h; that is, it is a sum of terms $c_{\alpha,\beta}\lambda^\alpha h^\beta$. Since all $\lambda \ge 0$ we can use $0 \le h_i < 1$, for all i, to bound the expression $c_{\alpha,\beta}\lambda^\alpha h^\beta$ by

$$0 \le c_{\alpha,\beta}\lambda^\alpha h^\beta < c_{\alpha,\beta}\lambda^\alpha$$

if $c_{\alpha,\beta} > 0$ and by

$$c_{\alpha,\beta}\lambda^\alpha < c_{\alpha,\beta}\lambda^\alpha h^\beta \le 0$$

if $c_{\alpha,\beta} < 0$. Putting now

$$g_l^{(j)}(\lambda) := \sum_{\alpha,\beta : c_{\alpha,\beta} < 0} c_{\alpha,\beta}\lambda^\alpha \quad \text{and} \quad g_u^{(j)}(\lambda) := \sum_{\alpha,\beta : c_{\alpha,\beta} > 0} c_{\alpha,\beta}\lambda^\alpha$$

we have

$$g_l^{(j)}(\lambda) \le v_x^{(j)} \le g_u^{(j)}(\lambda)$$

by construction. Moreover, again by construction, the degree of $g_l^{(j)}$ and of $g_u^{(j)}$ is strictly less than the degree of $g^{(j)}$. □

Acknowledgments. The authors wish to thank Jesus De Loera and an anonymous referee for many helpful comments. We express our thanks to Peter Malkin for pointing out [6].

REFERENCES

[1] D. Bertsimas and R. Weismantel, *Optimization Over Integers*, Dynamic Ideas, Belmont, MA, 2005.

[2] F. R. Giles and W. R. Pulleyblank, *Total dual integrality and integer polyhedra*, Linear Algebra Appl., 25 (1979), pp. 191–196.

[3] P. Gordan, *Über die Auflösung linearer Gleichungen mit reellen Coefficienten*, Math. Ann., 6 (1873), pp. 23–28.

[4] J. E. Graver, *On the foundation of linear and integer programming* I, Math. Programming, 9 (1975), pp. 207–226.

[5] U. U. Haus, M. Köppe, and R. Weismantel, *The integral basis method for integer programming*, Math. Methods Oper. Res., 53 (2001), pp. 281–307.

[6] R. G. Jeroslow, *Some basis theorems for integral monoids*, Math. Oper. Res., 3 (1978), pp. 145–154.

[7] A. Schrijver, *Theory of Linear and Integer Programming*, Wiley, Chichester, 1986.

[8] J. G. van der Corput, *Über Systeme von linear-homogenen Gleichungen und Ungleichungen*, Proceedings Koninklijke Akademie van Wetenschappen te Amsterdam, 34 (1931), pp. 368–371.

SIAM J. OPTIM.
Vol. 18, No. 1, pp. 138–155

MONGE PROPERTY AND BOUNDING MULTIVARIATE PROBABILITY DISTRIBUTION FUNCTIONS WITH GIVEN MARGINALS AND COVARIANCES*

XIAOLING HOU† AND ANDRÁS PRÉKOPA†

Abstract. Multivariate probability distributions with given marginals are considered, along with linear functionals, to be minimized or maximized, acting on them. The functionals are supposed to satisfy the Monge or inverse Monge or some higher order convexity property, and they may be only partially known. Existing results in connection with Monge arrays are reformulated and extended in terms of linear programming dual feasible bases. Lower and upper bounds are given for the optimum value as well as for unknown coefficients of the objective function, based on the knowledge of some dual feasible bases and corresponding objective function coefficients. In the two- and three-dimensional cases dual feasible bases are obtained for the problem, where not only the univariate marginals but also the covariances of the pairs of random variables are known.

Key words. distributions with given marginals, transportation problem, Monge arrays, bounding expectations under incomplete information

AMS subject classifications. 90C05, 90C08, 90C15, 60E05, 90B06

DOI. 10.1137/050638308

1. Introduction.

In this paper we consider multivariate (or multidimensional) discrete probability distributions with given marginals, along with special linear functionals, to be minimized or maximized, acting on them. In other words, we consider multidimensional transportation problems with special objective functions, where the sum of the marginal values is equal to 1. In the two- and three-dimensional cases we also look at distributions, where not only the marginals but also the covariances of the random variables involved are prescribed.

About the objective functions we assume that they enjoy the Monge or inverse Monge or some discrete higher order convexity property.

There is a considerable literature on the Monge property and its use in optimization and other fields of applied mathematics. The papers by Burkard, Klinz, and Rudolf [3] and Burkard [2] provide us with an overview of its history and the most important results. For basic results in connection with multivariate discrete higher order convexity, see Prékopa [6, 7] and Mádi-Nagy and Prékopa [5].

The purpose of this paper is the following. First, we reformulate the Monge and inverse Monge properties in terms of dual feasible bases of the transportation problem and obtain further results for them. Second, we give lower and upper bounds for the optimum value based on the knowledge of the univariate marginals and the covariances of pairs of bivariate marginals. The results for the latter case concern the two- and three-dimensional transportation problems. Third, we look at partially known objective functions and give lower and upper bounds for the unknown entries of the coefficient array. The bounds are based on the knowledge of the univariate marginals in the general d-dimensional case and on additional knowledge of the covariances in the two- and three-dimensional cases.

*Received by the editors August 17, 2005; accepted for publication (in revised form) July 18, 2006; published electronically February 23, 2007.
http://www.siam.org/journals/siopt/18-1/63830.html
†RUTCOR, Rutgers University, 640 Bartholomew Rd., Piscataway, NJ 08854 (xhou@rutcor.rutgers.edu, prekopa@rutcor.rutgers.edu). The work of the first author was supported by the DIMACS Winter 2004 and Summer 2005 Research Program.

The d-dimensional ($d \geq 2$) transportation problem is the following (linear program) LP:

$$
\begin{array}{ll}
\min(\max) & \sum_{i_1,\ldots,i_d} c(i_1,\ldots,i_d)x(i_1,\ldots,i_d) \\
\text{subject to} & \sum_{i_1,\ldots,i_d,i_k=i} x(i_1,\ldots,i_d) = a_k(i) \\
& \text{for all } \; i = 1,\ldots,n_k, \quad k = 1,\ldots,d, \\
& x(i_1,\ldots,i_d) \geq 0, \\
& \text{for all } \; i_k = 1,\ldots,n_k, \quad k = 1,\ldots,d.
\end{array} \tag{1}
$$

Let $\mathbf{A} = (\{\mathbf{a}(i_1,\ldots,i_d)\})$, $\mathbf{b} = (a_1(1),\ldots,a_1(n_1),\ldots,a_d(1),\ldots,a_d(n_d))^T$, $\mathbf{c} = (\{c(i_1,\ldots,i_d)\})$, $\mathbf{x} = (\{x(i_1,\ldots,i_d)\})$ be the coefficient matrix, the right-hand-side vector, the vector of the objective function coefficients, and the decision vector, respectively. Then problem (1) can be written as $\min(\max)\mathbf{c}^T\mathbf{x}$, subject to $\mathbf{A}\mathbf{x} = \mathbf{b}$, $\mathbf{x} \geq 0$. We assume that the above vectors and components are arranged in such a way that $(i_1,\ldots,i_d)$ follow the lexicographic order starting with the smallest. In what follows, the d-tuples $(i_1,\ldots,i_d)$ will be referred to as cells, and their collection will be called a transportation tableau.

If in problem (1) we have the relation $\sum_{i=1}^{n_1} a_1(i) = \cdots = \sum_{i=1}^{n_d} a_d(i) = 1$, then $\{x(i_1,\ldots,i_d)\}$ is a multivariate probability distribution, where its univariate marginals are prescribed. Problem (1) can then be reformulated in such a way that we minimize or maximize the expectation of $c(X_1,\ldots,X_d)$, where $X_1,\ldots,X_d$ are random variables with given distributions, $P(X_k = i) = a_k(i)$, $i = 1,\ldots,n_k$, $k = 1,\ldots,d$. Note that $\{1,\ldots,n_k\}$ serves as the support of X_k, $k = 1,\ldots,d$, but the solution of problem (1) does not depend on the choices of these sets.

In what follows we assume that the sum of each set of marginal values is equal to 1. Our results, however, generalize in a trivial way for the case where the sum of the marginal values is not 1.

A d-dimensional $n_1 \times \cdots \times n_d$ array $\mathbf{c} = \{c(i_1,\ldots,i_d)\}$ is called a *Monge array* if for all entries $c(i_1,\ldots,i_d)$ and $c(j_1,\ldots,j_d)$, $1 \leq i_k, j_k \leq n_k$, $1 \leq k \leq d$, we have

$$
c(s_1,\ldots,s_d) + c(t_1,\ldots,t_d) \leq c(i_1,\ldots,i_d) + c(j_1,\ldots,j_d), \tag{2}
$$

where for all $1 \leq k \leq d$, $s_k = \min\{i_k, j_k\}$, $t_k = \max\{i_k, j_k\}$. If (2) holds strictly for all $(s_1,\ldots,s_d) \neq (i_1,\ldots,i_d)$ and $(s_1,\ldots,s_d) \neq (j_1,\ldots,j_d)$, then we say that $\mathbf{c}$ is a *strict Monge array*. If the above inequalities are reversed, then $\mathbf{c}$ is called an *inverse Monge array* or a *strict inverse Monge array*.

To define discrete higher order convexity, let $Z_j = \{z_{j0},\ldots,z_{jn_j}\}$, $j = 1,\ldots,d$, be distinct element finite sets, and $f(z)$, $z \in Z = Z_1 \times \cdots \times Z_d$, a multivariate discrete function. Take a subset of Z,

$$
Z_{I_1,\ldots,I_d} = \{z_{1i}, i \in I_1\} \times \cdots \times \{z_{di}, i \in I_d\} = Z_{I_1} \times \cdots \times Z_{I_d},
$$

where $|I_j| = k_j + 1$, $k_j \leq n_j$, $j = 1,\ldots,d$. The $(k_1,\ldots,k_d)$*-order divided difference*, corresponding to $Z_{I_1,\ldots,I_d}$, will be designated by $[z_{1i}, i \in I_1; \ldots; z_{di}, i \in I_d; f]$. The sum $k_1 + \cdots + k_d$ is called its *total order*. We call the discrete function $(k_1,\ldots,k_d)$-order convex if all sequences in all $Z_1,\ldots,Z_d$ are increasing and all $(k_1,\ldots,k_d)$-order divided differences are nonnegative.

Our presentation is based, to a large extent, on linear programming theory. In this respect we use the notations and definitions presented in Prékopa [6].

We also introduce the concept of an ordered sequence for the columns of matrix $\mathbf{A}$ in problem (1). The collection of columns $\{\mathbf{a}(i_1,\ldots,i_d), (i_1,\ldots,i_d) \in I\}$ is called

an *ordered sequence* if I has the following form:

$$\begin{aligned} I = \{&(1,\dots,1,1),\dots,(1,\dots,1,i_{d,1}),\dots,(1,\dots,i_{d-1,1},i_{d,1}),\\ &\dots,(i_{1,1},\dots,i_{d-1,1},i_{d,1}),\dots,(i_{1,n-1},\dots,i_{d-1,n-1},i_{d,n-1}),\\ &\dots,(i_{1,n-1},\dots,i_{d-1,n-1},i_{d,n}),\dots,(i_{1,n-1},\dots,i_{d-1,n},i_{d,n}),\\ &\dots,(i_{1,n},\dots,i_{d-1,n},i_{d,n})\}, \end{aligned}$$

where $1 \le i_{k,1} \le \cdots \le i_{k,n-1} \le i_{k,n} = n_k$ for all $1 \le k \le d$. In what follows, the terms *ordered sequence* and *basis* for the columns of $\mathbf{A}$ will briefly be referred to as ordered *sequence* and *basis*, respectively. It is easy to see that the following assertion holds true: any ordered sequence forms a basis.

When $d = 2$ in problem (1), we call the collection of columns $\{\mathbf{a}_{ij},\ (i,j) \in J\}$ an *inverse ordered sequence* if J has the following form:

$$\begin{aligned} J = \{&(1,n-j_0),\dots,(1,n-j_1),(2,n-j_1),\dots,(2,n-j_2),\\ &\dots,(m,n-j_{m-1}),\dots,(m,n-j_m)\}, \end{aligned}$$

where $0 = j_0 \le j_1 \le j_2 \le \cdots \le j_{m-1} \le j_m = n-1$. The following assertion holds true: any inverse ordered sequence forms a basis.

An ordered (inverse ordered) sequence is a chain (antichain) in the partially ordered set of the cells.

The further parts of the paper are organized as follows. Section 2 is devoted to the study of the bivariate and multivariate cases. Existing results in connection with Monge and distribution arrays are reformulated and extended in terms of dual feasibility of bases. Bounds on the expectation and the unknown components of $\mathbf{c}$ are obtained under the condition that the univariate marginals of the random vector are known. In section 3 a new algorithm called $(GREEDY\ DUAL)_d$ for solving the d-dimensional transportation problem is presented. In section 4 the bivariate case is considered, where, in addition to the knowledge of the marginal distributions, we assume the knowledge of the covariance of the two random variables involved. We give bounds for the same values as before. In section 5 we present similar results for the three-dimensional case. Finally, some applications and illustrative examples are given in section 6.

2. Monge property and dual feasible bases. In this section we establish relationship between ordered sequences and dual feasible bases in the d-dimensional problem (1), and between the inverse ordered sequences and the dual feasible bases in the two-dimensional problem (1).

First, we present the relationship between ordered sequences and the dual feasible bases in the d-dimensional problem (1). To prove our results we recall three theorems from linear programming. For a proof of the first one, see Prékopa [6, Theorem 5]; the second one can be derived from Theorem 3 of the same paper.

THEOREM 2.1. *If problem* (1) *has a primal feasible solution and a finite optimum, then there exists a primal feasible basis that is also dual feasible.*

THEOREM 2.2. *If in problem* (1) B *is a nondegenerate optimal basis, then* $\mathbf{B}$ *is dual feasible.*

Bein et al. [1] extended the two-dimensional greedy algorithm $GREEDY_2$, due to Hoffman [4], to the d-dimensional greedy algorithm $GREEDY_d$ and proved the following theorem.

THEOREM 2.3 (see Bein et al. [1]). *Given a particular d-dimensional* $n_1 \times \cdots \times n_d$ *cost array* $\mathbf{c}$*, the algorithm* $GREEDY_d$ *solves the corresponding d-dimensional transportation problem for any* $\mathbf{b}$ *if and only if* $\mathbf{c}$ *is Monge.*

For the d-dimensional minimization problem (1), we prove two theorems.

THEOREM 2.4. *In the minimization problem* (1)*, any ordered sequence forms a dual feasible basis if and only if* $\mathbf{c}$ *satisfies the Monge property.*

Proof. For the proof of the "if" direction, assume that $\mathbf{c}$ is Monge. For any given ordered sequence write positive numbers in its cells, and call what comes out on the right-hand side (r.h.s.) $\mathbf{b}$. For this $\mathbf{b}$ the algorithm $GREEDY_d$ produces the same ordered sequence. By Theorem 2.3 this ordered sequence is a primal nondegenerate optimal basis to the problem. By Theorem 2.2, this ordered sequence is dual feasible.

For the proof of the "only if" direction, assume that any ordered sequence forms a dual feasible basis. Then for any $\mathbf{b}$ the algorithm $GREEDY_d$ solves the problem optimally, because the algorithm $GREEDY_d$ produces an ordered sequence. Thus, by Theorem 2.3, $\mathbf{c}$ is Monge. □

To prove the next theorem we need the dual of the minimization problem (1) that is given as follows:

$$\begin{aligned} \max \quad & \textstyle\sum_{k=1}^{d} \sum_{i_k=1}^{n_k} a_k(i_k) w_k(i_k) \\ \text{subject to} \quad & \\ & w_1(i_1) + \cdots + w_d(i_d) \le c(i_1, \ldots, i_d) \\ & \text{for all} \quad i_k = 1, \ldots, n_k, \quad k = 1, \ldots, d. \end{aligned}$$

THEOREM 2.5. *If, in the minimization problem* (1)*,* $\mathbf{c}$ *satisfies the strict Monge property, then any dual feasible basis is dual nondegenerate and forms an ordered sequence.*

Proof. Suppose that $\mathbf{c}$ satisfies the strict Monge property and that $\mathbf{B}$ is a dual feasible basis. First, assume that the vectors of $\mathbf{B}$ do not form an ordered sequence. Then there must exist vectors $\mathbf{a}(i_1, \ldots, i_d)$ and $\mathbf{a}(j_1, \ldots, j_d)$ in $\mathbf{B}$ such that if $s_k = \min\{i_k, j_k\}$, $t_k = \max\{i_k, j_k\}$, then $(s_1, \ldots, s_d) \neq (i_1, \ldots, i_d)$, $(s_1, \ldots, s_d) \neq (j_1, \ldots, j_d)$. Let $w_k(i_k)$, $i_k = 1, \ldots, n_k$, $k = 1, \ldots, d$, be the components of the dual vector corresponding to $\mathbf{B}$. Then

$$\begin{aligned} c(i_1, \ldots, i_d) &= w_1(i_1) + \cdots + w_d(i_d), \\ c(j_1, \ldots, j_d) &= w_1(j_1) + \cdots + w_d(j_d), \\ c(s_1, \ldots, s_d) &\ge w_1(s_1) + \cdots + w_d(s_d), \\ c(t_1, \ldots, t_d) &\ge w_1(t_1) + \cdots + w_d(t_d). \end{aligned} \tag{3}$$

Since

$$\begin{aligned} w_1(i_1) + \cdots + w_d(i_d) + w_1(j_1) + \cdots + w_d(j_d) & \\ = w_1(s_1) + \cdots + w_d(s_d) + w_1(t_1) + \cdots + w_d(t_d), & \end{aligned} \tag{4}$$

we have the relation

$$c(s_1, \ldots, s_d) + c(t_1, \ldots, t_d) \ge c(i_1, \ldots, i_d) + c(j_1, \ldots, j_d).$$

This contradicts the strict Monge property, so any dual feasible basis must be an ordered sequence.

TABLE 1
"min (max)" means minimization (maximization) problem (1)*, "o.s." means ordered sequence, and "d.f.b." means dual feasible basis.*

	d-dimensional min (max)
Any o.s. forms a d.f.b.	*iff* **c** is (inverse) Monge
Any d.f.b. forms an o.s.	*if* **c** is strict (inverse) Monge

TABLE 2
"i.o.s." means inverse ordered sequence.

	two-dimensional max (min)
Any i.o.s. forms a d.f.b.	*iff* **c** is (inverse) Monge
Any d.f.b. forms an i.o.s.	*if* **c** is strict (inverse) Monge

Second, suppose that the vector $\mathbf{a}(i_1, \ldots, i_d)$ is nonbasic. Since $\mathbf{B}$ forms an ordered sequence, we can find a basic vector $\mathbf{a}(j_1, \ldots, j_d)$ in $\mathbf{B}$ such that if $s_k = \min\{i_k, j_k\}$, $t_k = \max\{i_k, j_k\}$, then $(s_1, \ldots, s_d) \neq (i_1, \ldots, i_d)$, $(s_1, \ldots, s_d) \neq (j_1, \ldots, j_d)$. It is easy to see that the last three equations in (3) and the one in (4) hold true. Because of the strict Monge property, we have the following relations:

$$\begin{aligned} c(i_1, \ldots, i_d) + c(j_1, \ldots, j_d) &> c(s_1, \ldots, s_d) + c(t_1, \ldots, t_d) \\ &\geq w_1(s_1) + \cdots + w_d(s_d) + w_1(t_1) + \cdots + w_d(t_d) \\ &= w_1(i_1) + \cdots + w_d(i_d) + w_1(j_1) + \cdots + w_d(j_d). \end{aligned}$$

So

$$c(i_1, \ldots, i_d) > w_1(i_1) + \cdots + w_d(i_d).$$

Therefore $\mathbf{B}$ is dual nondegenerate. □

From these two theorems, we can easily get the similar results for the d-dimensional maximization problem (1). We summarize all these results in Table 1.

In the two-dimensional case let $c'_{ij} = c_{(m-i+1)j}$, $\mathbf{a}'_{ij} = \mathbf{a}_{(m-i+1)j}$. Then $\mathbf{c} = (c_{ij})$ is an inverse Monge array if and only if $\mathbf{c}' = (c'_{ij}) = (c_{(m-i+1)j})$ is a Monge array. $\{\mathbf{a}_{ij},\ (i,j) \in J\}$ is an inverse ordered sequence of $\mathbf{A}$ if and only if $\{\mathbf{a}'_{ij},\ (i,j) \in J\}$ is an ordered sequence of $\mathbf{A}' = (\mathbf{a}'_{ij})$. Thus, for the two-dimensional problem (1) we obtain the relationship between the inverse ordered sequences and the dual feasible bases given in Table 2.

If $\mathbf{c}$ is Monge in the three-dimensional maximization problem (1), then we have the following theorem.

THEOREM 2.6. *Consider the three-dimensional maximization problem* (1)*. If* $\mathbf{c}$ *is Monge, then each of the following sequences of vectors forms a dual feasible basis:*

$$\begin{aligned} (S'_1)\quad & \{\mathbf{a}_{i11}, i = 1, \ldots, n_1, \mathbf{a}_{1j1}, j = 2, \ldots, n_2, \mathbf{a}_{11k}, k = 2, \ldots, n_3\}, \\ (S'_2)\quad & \{\mathbf{a}_{in_2n_3}, i = 1, \ldots, n_1, \mathbf{a}_{n_1jn_3}, j = 1, \ldots, n_2 - 1, \\ & \quad \mathbf{a}_{n_1n_2k}, k = 1, \ldots, n_3 - 1\}. \end{aligned}$$

3. The (*GREEDY DUAL*)$_d$ algorithm. In addition to the $GREEDY_d$ algorithm a dual algorithm can also be constructed to solve the d-dimensional transportation problem with a Monge array in the objective function.

Consider an ordered sequence, and single out from it those cells which are "turning points" in the sequence of cells. Let us call them *pivot cells*. Thus, if in the ordered

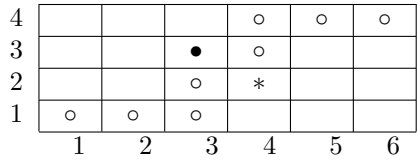


FIG. 1. *"○" means basic cell, "●" means the leaving (entering) cell, and "∗" means the entering (leaving) cell.*

sequence we have one of the following consecutive cells,

$$(\dots, i-1, \dots, j, \dots)\ (\dots, i, \dots, j, \dots)\ (\dots, i, \dots, j+1, \dots), \tag{5}$$

$$(\dots, i, \dots, j-1, \dots)\ (\dots, i, \dots, j, \dots)\ (\dots, i+1, \dots, j, \dots), \tag{6}$$

where the elements represented by dots remains unchanged, then the cell $(\dots, i, \dots, j, \dots)$ is a pivot cell. The algorithm can be described as follows.

Step 0. Choose arbitrarily an ordered sequence. By Theorem 2.4 it represents a dual feasible basis. Compute the corresponding basic solution.

Step 1. Check whether the basic components of the basic solution are nonnegative. In view of the constraints in problem (1) only pivot cells may contain negative entries; thus it is enough to check the entries in the pivot cells. If all pivot entries are nonnegative, then stop; optimal basis and optimal solution have been found.

Step 2. Choose arbitrarily a pivot cell that contains a negative entry, and let it leave the basis.

Step 3. If the leaving cell is the second one in (5) (in (6)), then let the cell $(\dots, i-1, \dots, j+1, \dots)$ (the cell $(\dots, i+1, \dots, j-1, \dots)$) enter the basis.

Step 4. Compute the basic solution $\mathbf{x}'$ corresponding to the new basis. In the case of (5) we have

$$x'_{(\dots,i-1,\dots,j,\dots)} = x_{(\dots,i-1,\dots,j,\dots)} + x_{(\dots,i,\dots,j,\dots)} x'_{(\dots,i,\dots,j,\dots)} = 0,$$
$$x'_{(\dots,i,\dots,j+1,\dots)} = x_{(\dots,i,\dots,j+1,\dots)} + x_{(\dots,i,\dots,j,\dots)} x'_{(\dots,i-1,\dots,j+1,\dots)} = -x_{(\dots,i,\dots,j,\dots)}.$$

All other $\mathbf{x}'$ components remain unchanged. The basic solution for the case of (6) can be obtained similarly. Go to Step 1.

Figure 1 provides an illustration of the leaving and entering cells.

With straightforward modification in the above algorithm we can create one to solve the problem with an inverse Monge array.

The above algorithm does not depend on $\mathbf{c}$ as long as it is Monge. Since all dual feasible bases are nondegenerate if $\mathbf{c}$ has the strict Monge property, it follows that the algorithm terminates in a finite number of steps.

4. The use of covariance in the two-dimensional case. In this section we consider the following problem:

$$\begin{aligned} &\min(\max) && \textstyle\sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}x_{ij} \\ &\text{subject to} \\ &&& \textstyle\sum_{j=1}^{n} x_{ij} = a_i, \quad i = 1, \dots, m, \\ &&& \textstyle\sum_{i=1}^{m} x_{ij} = b_j, \quad j = 1, \dots, n, \\ &&& \textstyle\sum_{i=1}^{m}\sum_{j=1}^{n} y_i z_j x_{ij} = d, \\ &&& x_{ij} \geq 0, \quad i = 1, \dots, m, \quad j = 1, \dots, n. \end{aligned} \tag{7}$$

We have in mind a pair of discrete random variables, Y, Z, with supports $\{y_i\}$, $\{z_j\}$, respectively, where the univariate marginal distributions as well as $E(YZ)$ are prescribed but the joint distribution of Y and Z is not specified otherwise. Since the marginal distributions determine $E(Y)$ and $E(Z)$, to prescribe $E(YZ)$ is the same as to prescribe $Cov(Y,Z) = E(YZ) - E(Y)E(Z)$. Since we can write c_{ij} in the form $c_{ij} = c(y_i, z_j)$, it follows that c is a function of the random vector (y, z), and if we plug in the random variables Y, Z, it becomes a function of (Y, Z). Thus, the optimum values of problem (7) provide us with the best possible lower and upper bounds for $E(c(Y,Z))$ under the given conditions.

Problem (7) can be written in the following matrix form: $\min(\max)\mathbf{c}^T\mathbf{x}$, subject to $\overline{\mathbf{A}}\mathbf{x} = \overline{\mathbf{b}}$, $\mathbf{x} \geq 0$, where $\overline{\mathbf{A}} = (\overline{\mathbf{a}}_{ij})$, $\overline{\mathbf{a}}_{ij} = \mathbf{e}_i + \mathbf{e}_{m+j} + y_i z_j \mathbf{e}_{m+n+1}$, $i = 1, \ldots, m$, $j = 1, \ldots, n$; $\mathbf{e}_i$, $\mathbf{e}_{m+j}$, and $\mathbf{e}_{m+n+1}$ are unit vectors in $\mathbf{E}^{m+n+1}$ with ones in the ith, $(m+j)$th, and $(m+n+1)$th positions, respectively; and $\overline{\mathbf{b}} = (\mathbf{b}^T, d)^T$.

THEOREM 4.1. *Consider the minimization problem* (7)*, and assume that* $\{y_i\}$ *is strictly increasing,* $\{z_j\}$ *is strictly decreasing, and the* (1,2)*-order and* (2,1)*-order divided differences of* $c(y_i, z_j)$ *are nonnegative. Then* $B_1 = \{\overline{\mathbf{a}}_{i1}, i = 1, \ldots, m, \overline{\mathbf{a}}_{mj}, j = 2, \ldots, n, \overline{\mathbf{a}}_{1n}\}$ *forms a dual feasible basis of the columns of* $\overline{\mathbf{A}}$*.*

Proof. First, let us show that B_1 forms a basis of the columns of $\overline{\mathbf{A}}$. It is easy to see that the rank of $\overline{\mathbf{A}}$ is $m+n$, and there are $m+n$ vectors in B_1. To show the linear independence of the $m+n$ vectors in B_1, consider the linear combination of the vectors of B_1:

$$\begin{aligned}
&\sum_{i=1}^{m} \lambda_{i1}\overline{\mathbf{a}}_{i1} + \sum_{j=2}^{n} \lambda_{mj}\overline{\mathbf{a}}_{mj} + \lambda_{1n}\overline{\mathbf{a}}_{1n} \\
&\quad = \sum_{i=1}^{m} \lambda_{i1}(\mathbf{e}_i + \mathbf{e}_{m+1} + y_i z_1 \mathbf{e}_{m+n+1}) + \sum_{j=2}^{n} \lambda_{mj}(\mathbf{e}_m + \mathbf{e}_{m+j} + y_m z_j \mathbf{e}_{m+n+1}) \\
&\qquad + \lambda_{1n}(\mathbf{e}_1 + \mathbf{e}_{m+n} + y_1 z_n \mathbf{e}_{m+n+1}) \\
&\quad = (\lambda_{11} + \lambda_{1n})\mathbf{e}_1 + \sum_{i=2}^{m-1} \lambda_{i1}\mathbf{e}_i + \left(\sum_{j=1}^{n} \lambda_{mj}\right)\mathbf{e}_m + \left(\sum_{i=1}^{m} \lambda_{i1}\right)\mathbf{e}_{m+1} + \sum_{j=2}^{n-1} \lambda_{mj}\mathbf{e}_{m+j} \\
&\qquad + (\lambda_{mn} + \lambda_{1n})\mathbf{e}_{m+n} + \left(\sum_{i=1}^{m} \lambda_{i1} y_i z_1 + \sum_{j=2}^{n} \lambda_{mj} y_m z_j + \lambda_{1n} y_1 z_n\right)\mathbf{e}_{m+n+1}.
\end{aligned}$$

If it equals 0, then, by the linear independence of the unit vectors, it follows that all λ's must be 0.

Secondly, let us show that this basis is dual feasible. For any nonbasic vector $\overline{\mathbf{a}}_{ij}$, $1 \leq i < m$, $1 < j \leq n$, we have the equations

$$\begin{aligned}
\overline{\mathbf{a}}_{ij} - \overline{\mathbf{a}}_{i1} + \overline{\mathbf{a}}_{m1} - \overline{\mathbf{a}}_{mj} &= (y_i - y_m)(z_j - z_1)\mathbf{e}_{m+n+1} \\
\overline{\mathbf{a}}_{ij} - \overline{\mathbf{a}}_{i1} + \overline{\mathbf{a}}_{11} - \overline{\mathbf{a}}_{1n} + \overline{\mathbf{a}}_{mn} - \overline{\mathbf{a}}_{mj} & \\
= [(y_i - y_m)(z_j - z_1) - (y_1 - y_m)(z_n - z_1)]\mathbf{e}_{m+n+1}. &
\end{aligned}$$

From here we derive the expression of $\overline{\mathbf{a}}_{\mathbf{ij}}$ as the following linear combination of the basic vectors:

$$\begin{aligned}
\overline{\mathbf{a}}_{ij} = &\frac{(y_1 - y_m)(z_1 - z_n) - (y_i - y_m)(z_1 - z_j)}{(y_1 - y_m)(z_1 - z_n)}(\overline{\mathbf{a}}_{i1} - \overline{\mathbf{a}}_{m1} + \overline{\mathbf{a}}_{mj}) \\
&- \frac{(y_i - y_m)(z_1 - z_j)}{(y_1 - y_m)(z_1 - z_n)}(\overline{\mathbf{a}}_{11} - \overline{\mathbf{a}}_{1n} + \overline{\mathbf{a}}_{mn} - \overline{\mathbf{a}}_{mj} - \overline{\mathbf{a}}_{i1}).
\end{aligned}$$

We have to prove that

$$\frac{(y_1-y_m)(z_1-z_n)-(y_i-y_m)(z_1-z_j)}{(y_1-y_m)(z_1-z_n)}(c_{i1}-c_{m1}+c_{mj})$$
$$-\frac{(y_i-y_m)(z_1-z_j)}{(y_1-y_m)(z_1-z_n)}(c_{11}-c_{1n}+c_{mn}-c_{mj}-c_{i1})-c_{ij}\le 0,$$

or, equivalently,

$$(c_{i1}-c_{m1}+c_{mj}-c_{ij})\le\frac{(y_i-y_m)(z_1-z_j)}{(y_1-y_m)(z_1-z_n)}(c_{11}-c_{m1}+c_{mn}-c_{1n}). \tag{8}$$

Since $(y_i-y_m)(z_1-z_j)<0$, the above inequality is equivalent to

$$\frac{c_{i1}-c_{m1}-c_{ij}+c_{mj}}{(y_i-y_m)(z_1-z_j)}\ge\frac{c_{11}-c_{m1}-c_{1n}+c_{mn}}{(y_1-y_m)(z_1-z_n)}.$$

We have assumed that the (1,2)-order and (2,1)-order divided differences of $c(y_i,z_j)$ are nonnegative. The nonnegativity of the (1,2)-order divided difference implies

$$\frac{\frac{c_{11}-c_{m1}-c_{1n}+c_{mn}}{(y_1-y_m)(z_1-z_n)}-\frac{c_{11}-c_{m1}-c_{1j}+c_{mj}}{(y_1-y_m)(z_1-z_j)}}{z_n-z_j}\ge 0.$$

Similarly, the nonnegativity of (2,1)-order divided difference gives

$$\frac{\frac{c_{11}-c_{m1}-c_{1j}+c_{mj}}{(y_1-y_m)(z_1-z_j)}-\frac{c_{i1}-c_{m1}-c_{ij}+c_{mj}}{(y_i-y_m)(z_1-z_j)}}{y_1-y_i}\ge 0.$$

Since both z_n-z_j and y_1-y_i are negative, the above two inequalities imply

$$\frac{c_{i1}-c_{m1}-c_{ij}+c_{mj}}{(y_i-y_m)(z_1-z_j)}\ge\frac{c_{11}-c_{m1}-c_{1j}+c_{mj}}{(y_1-y_m)(z_1-z_j)}\ge\frac{c_{11}-c_{m1}-c_{1n}+c_{mn}}{(y_1-y_m)(z_1-z_n)}.$$

This completes the proof. □

If we use the reasoning in the proof of Theorem 4.1, we can obtain a variety of dual feasible bases for problem (7), under different conditions. Let $B_2=\{\bar{\mathbf{a}}_{in}, i=1,\dots,m,\bar{\mathbf{a}}_{1j}, j=1,\dots,n-1,\bar{\mathbf{a}}_{m1}\}$, $B_3=\{\bar{\mathbf{a}}_{i1}, i=1,\dots,m,\bar{\mathbf{a}}_{1j}, j=2,\dots,n,\bar{\mathbf{a}}_{mn}\}$, $B_4=\{\bar{\mathbf{a}}_{in}, i=1,\dots,m,\bar{\mathbf{a}}_{mj}, j=1,\dots,n-1,\bar{\mathbf{a}}_{11}\}$. The cells corresponding to the vectors in B_1, B_2, B_3 and B_4 are designated by **boldface** points in Figure 2.

We summarize all the results for the minimization (maximization) problem (7) in Table 3.

5. The use of covariances in the three-dimensional case. In this section we consider the following three-dimensional problem:

$$\begin{array}{ll}
\min(\max) & \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3} c_{ijk}x_{ijk}\\
\text{subject to} & \\
& \sum_{j=1}^{n_2}\sum_{k=1}^{n_3} x_{ijk}=a_i,\quad i=1,\dots,n_1,\\
& \sum_{i=1}^{n_1}\sum_{k=1}^{n_3} x_{ijk}=b_j,\quad j=1,\dots,n_2,\\
& \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} x_{ijk}=c_k,\quad k=1,\dots,n_3,\\
& \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3} y_iz_jx_{ijk}=d_1,\\
& \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3} z_jw_kx_{ijk}=d_2,\\
& \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3} y_iw_kx_{ijk}=d_3,\\
& x_{ijk}\ge 0,\quad i=1,\dots,n_1,\quad j=1,\dots,n_2,\quad k=1,\dots,n_3.
\end{array} \tag{9}$$

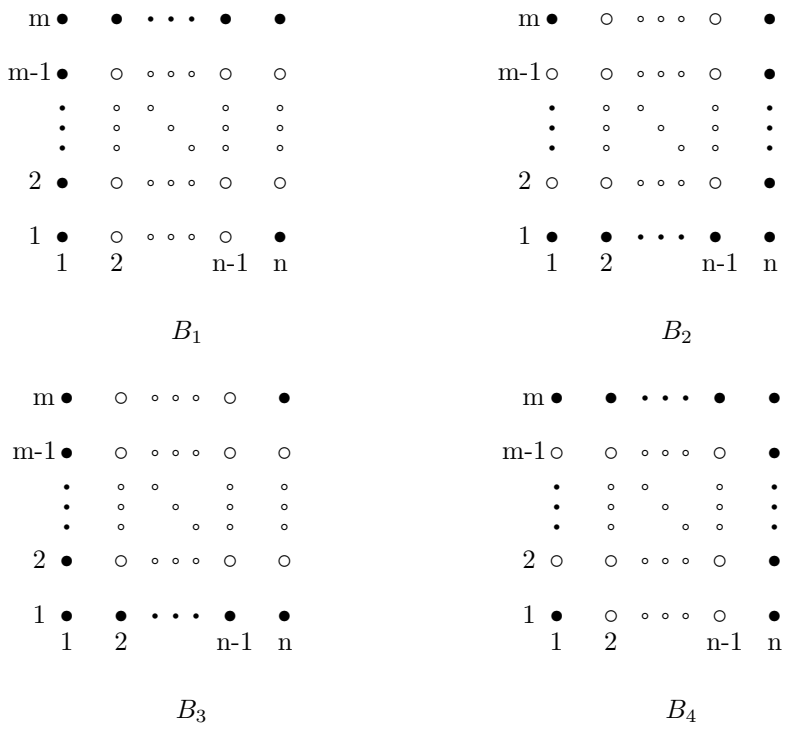


FIG. 2.

TABLE 3
"↗" means strictly increasing, "↘" means strictly decreasing. "(i,j)-order" means (i,j)-order divided difference of c, $i,j=1,2$, and "d.f.b. in min (max)" means dual feasible basis in the minimization (maximization) problem (7).

y_i	z_j	(1,2)-order	(2,1)-order	d.f.b. in min	d.f.b. in max
↗	↘	≥ 0	≥ 0	B_1	B_2
↘	↗	≤ 0	≤ 0	B_1	B_2
↗	↗	≥ 0	≤ 0	B_1	B_2
↘	↘	≤ 0	≥ 0	B_1	B_2
↗	↘	≤ 0	≤ 0	B_2	B_1
↘	↗	≥ 0	≥ 0	B_2	B_1
↗	↗	≤ 0	≥ 0	B_2	B_1
↘	↘	≥ 0	≤ 0	B_2	B_1
↗	↘	≤ 0	≥ 0	B_3	B_4
↘	↗	≥ 0	≤ 0	B_3	B_4
↗	↗	≤ 0	≤ 0	B_3	B_4
↘	↘	≥ 0	≥ 0	B_3	B_4
↗	↘	≥ 0	≤ 0	B_4	B_3
↘	↗	≤ 0	≥ 0	B_4	B_3
↗	↗	≥ 0	≥ 0	B_4	B_3
↘	↘	≤ 0	≤ 0	B_4	B_3

This problem can be interpreted in such a way that, given the supports $\{y_i\}$, $\{z_j\}$, $\{w_k\}$, the univariate marginals $\{a_i\}$, $\{b_j\}$, $\{c_k\}$, and the covariances of all pairs of three random variables Y, Z, W, we want to find the best possible lower and upper bounds for $E(c(Y,Z,W))$, where $c(y_i,z_j,w_k)=c_{ijk}$ for all i,j,k. Problem (9) can be written in the compact form $\min(\max)\mathbf{c}^T\mathbf{x}$, subject to $\overline{\mathbf{A}}\mathbf{x}=\overline{\mathbf{b}}$, $\mathbf{x}\geq 0$, where

$\overline{\mathbf{A}} = (\overline{\mathbf{a}}_{ijk})$, $\overline{\mathbf{a}}_{ijk} = \mathbf{e}_i + \mathbf{e}_{n_1+j} + \mathbf{e}_{n_1+n_2+k} + y_i z_j \mathbf{e}_{n_1+n_2+n_3+1} + z_j w_k \mathbf{e}_{n_1+n_2+n_3+2} + y_i w_k \mathbf{e}_{n_1+n_2+n_3+3}$, $i = 1, \ldots, n_1$, $j = 1, \ldots, n_2$, $k = 1, \ldots, n_3$; $\mathbf{e}_l$, $1 \le l \le n_1+n_1+n_3+3$, is the unit vector in $\mathbf{E}^{n_1+n_2+n_3+3}$ with one in the lth position; and $\overline{\mathbf{b}} = (\mathbf{b}^T, \mathbf{d}^T)^T$, $\mathbf{d} = (d_1, d_2, d_3)^T$.

Let $\overline{S}_1 = \{\overline{\mathbf{a}}_{i11}, i = 1, \ldots, n_1, \overline{\mathbf{a}}_{n_1 j 1}, j = 2, \ldots, n_2, \overline{\mathbf{a}}_{n_1 n_2 k}, k = 2, \ldots, n_3\}$. We prove the following result.

THEOREM 5.1. *Consider the minimization problem* (9). *If y_i, z_j, and w_k are strictly increasing; all $(1,2,0)$-order, $(1,0,2)$-order, $(0,1,2)$-order divided differences of $c(y_i, z_j, w_k)$ are nonnegative; and all $(2,1,0)$-order, $(2,0,1)$-order, $(0,2,1)$-order, and $(1,1,1)$-order divided differences of $c(y_i, z_j, w_k)$ are nonpositive, then $\overline{S}_1 \cup \{\overline{\mathbf{a}}_{n_1 1 n_3}, \overline{\mathbf{a}}_{1 n_2 1}, \overline{\mathbf{a}}_{11 n_3}\}$ forms a dual feasible basis of the columns of $\overline{\mathbf{A}}$.*

Proof. Similarly with Theorem 4.1 we can show that the vectors of the sequence $\overline{S}_1$ supplemented by the vectors $\overline{\mathbf{a}}_{n_1 1 n_3}$, $\overline{\mathbf{a}}_{1 n_2 1}$, and $\overline{\mathbf{a}}_{11 n_3}$ form a basis of the columns of $\overline{\mathbf{A}}$.

Now let us show that this basis is dual feasible. Assume that all $\{y_i\}$, $\{z_j\}$, and $\{w_k\}$ sequences are strictly increasing. For any nonbasic vector $\overline{\mathbf{a}}_{ijk}$, $1 \le i \le n_1$, $1 \le j \le n_2$, $1 \le k \le n_3$, we have the following four equations:

$$\begin{aligned} &\overline{\mathbf{a}}_{ijk} - \overline{\mathbf{a}}_{i11} + \overline{\mathbf{a}}_{n_1 11} - \overline{\mathbf{a}}_{n_1 j 1} + \overline{\mathbf{a}}_{n_1 n_2 1} - \overline{\mathbf{a}}_{n_1 n_2 k} \\ &\quad = \begin{pmatrix} \mathbf{0} \\ (y_i - y_{n_1})(z_j - z_1) \\ (z_j - z_{n_2})(w_k - w_1) \\ (y_i - y_{n_1})(w_k - w_1) \end{pmatrix}, \end{aligned} \tag{10}$$

$$\begin{aligned} &\overline{\mathbf{a}}_{ijk} - \overline{\mathbf{a}}_{i11} + \overline{\mathbf{a}}_{n_1 11} - \overline{\mathbf{a}}_{n_1 n_2 k} + \overline{\mathbf{a}}_{n_1 n_2 n_3} - \overline{\mathbf{a}}_{n_1 1 n_3} + \overline{\mathbf{a}}_{n_1 11} - \overline{\mathbf{a}}_{n_1 j 1} \\ &\quad = \begin{pmatrix} \mathbf{0} \\ (y_i - y_{n_1})(z_j - z_1) \\ (z_j - z_{n_2})(w_k - w_1) - (z_1 - z_{n_2})(w_{n_3} - w_1) \\ (y_i - y_{n_1})(w_k - w_1) \end{pmatrix}, \end{aligned} \tag{11}$$

$$\begin{aligned} &\overline{\mathbf{a}}_{ijk} - \overline{\mathbf{a}}_{n_1 j 1} + \overline{\mathbf{a}}_{n_1 n_2 1} - \overline{\mathbf{a}}_{n_1 n_2 k} + \overline{\mathbf{a}}_{n_1 1 n_3} - \overline{\mathbf{a}}_{11 n_3} + \overline{\mathbf{a}}_{111} - \overline{\mathbf{a}}_{i11} \\ &\quad = \begin{pmatrix} \mathbf{0} \\ (y_i - y_{n_1})(z_j - z_1) \\ (z_j - z_{n_2})(w_k - w_1) \\ (y_i - y_{n_1})(w_k - w_1) - (y_1 - y_{n_1})(w_{n_3} - w_1) \end{pmatrix}, \end{aligned} \tag{12}$$

$$\begin{aligned} &\overline{\mathbf{a}}_{ijk} - \overline{\mathbf{a}}_{i11} + \overline{\mathbf{a}}_{111} - \overline{\mathbf{a}}_{1 n_2 1} + \overline{\mathbf{a}}_{n_1 n_2 1} - \overline{\mathbf{a}}_{n_1 j 1} + \overline{\mathbf{a}}_{n_1 n_2 1} - \overline{\mathbf{a}}_{n_1 n_2 k} \\ &\quad = \begin{pmatrix} \mathbf{0} \\ (y_i - y_{n_1})(z_j - z_1) - (y_1 - y_{n_1})(z_{n_2} - z_1) \\ (z_j - z_{n_2})(w_k - w_1) \\ (y_i - y_{n_1})(w_k - w_1) \end{pmatrix}, \end{aligned} \tag{13}$$

where $\mathbf{0}$ is a zero vector in $\mathbf{R}^{n_1+n_2+n_3}$. For simplicity, let

$$\begin{aligned} A_1 &= (y_i - y_{n_1})(z_j - z_1), & A_2 &= (z_j - z_{n_2})(w_k - w_1), \\ A_3 &= (y_i - y_{n_1})(w_k - w_1), & B_1 &= (y_1 - y_{n_1})(z_{n_2} - z_1), \\ B_2 &= (z_1 - z_{n_2})(w_{n_3} - w_1), & B_3 &= (y_1 - y_{n_1})(w_{n_3} - w_1), \end{aligned}$$

$$\mathbf{a}_{12} = \begin{pmatrix} \mathbf{0} \\ A_1 \\ A_2 \\ 0 \end{pmatrix}, \quad \mathbf{a}_{13} = \begin{pmatrix} \mathbf{0} \\ A_1 \\ 0 \\ A_3 \end{pmatrix}, \quad \mathbf{a}_{23} = \begin{pmatrix} \mathbf{0} \\ 0 \\ A_2 \\ A_3 \end{pmatrix}.$$

From (10) and (11) we obtain

$$\frac{A_2 - B_2}{A_2}(\overline{\mathbf{a}}_{ijk} - \overline{\mathbf{a}}_{i11} + \overline{\mathbf{a}}_{n_1 11} - \overline{\mathbf{a}}_{n_1 j1} + \overline{\mathbf{a}}_{n_1 n_2 1} - \overline{\mathbf{a}}_{n_1 n_2 k} - \mathbf{a}_{13})$$
$$= \overline{\mathbf{a}}_{ijk} - \overline{\mathbf{a}}_{i11} + \overline{\mathbf{a}}_{n_1 11} - \overline{\mathbf{a}}_{n_1 n_2 k} + \overline{\mathbf{a}}_{n_1 n_2 n_3} - \overline{\mathbf{a}}_{n_1 1 n_3} + \overline{\mathbf{a}}_{n_1 11} - \overline{\mathbf{a}}_{n_1 j1} - \mathbf{a}_{13}.$$

Thus,

$$\begin{aligned}\overline{\mathbf{a}}_{ijk} &= (\overline{\mathbf{a}}_{i11} - \overline{\mathbf{a}}_{n_1 11} + \overline{\mathbf{a}}_{n_1 j1} - \overline{\mathbf{a}}_{n_1 n_2 1} + \overline{\mathbf{a}}_{n_1 n_2 k} + \mathbf{a}_{13}) \\ &\quad - \frac{A_2}{B_2}(\overline{\mathbf{a}}_{n_1 n_2 n_3} - \overline{\mathbf{a}}_{n_1 1 n_3} + \overline{\mathbf{a}}_{n_1 11} - \overline{\mathbf{a}}_{n_1 n_2 1}). \qquad (14)\end{aligned}$$

Equations (10) and (12) imply

$$\begin{aligned}\overline{\mathbf{a}}_{ijk} &= (\overline{\mathbf{a}}_{i11} - \overline{\mathbf{a}}_{n_1 11} + \overline{\mathbf{a}}_{n_1 j1} - \overline{\mathbf{a}}_{n_1 n_2 1} + \overline{\mathbf{a}}_{n_1 n_2 k} + \mathbf{a}_{12}) \\ &\quad - \frac{A_3}{B_3}(\overline{\mathbf{a}}_{n_1 1 n_3} - \overline{\mathbf{a}}_{11 n_3} + \overline{\mathbf{a}}_{111} - \overline{\mathbf{a}}_{n_1 11}). \qquad (15)\end{aligned}$$

Similarly, equations (10) and (13) imply

$$\begin{aligned}\overline{\mathbf{a}}_{ijk} &= (\overline{\mathbf{a}}_{i11} - \overline{\mathbf{a}}_{n_1 11} + \overline{\mathbf{a}}_{n_1 j1} - \overline{\mathbf{a}}_{n_1 n_2 1} + \overline{\mathbf{a}}_{n_1 n_2 k} + \mathbf{a}_{23}) \\ &\quad - \frac{A_1}{B_1}(\overline{\mathbf{a}}_{111} - \overline{\mathbf{a}}_{1 n_2 1} + \overline{\mathbf{a}}_{n_1 n_2 1} - \overline{\mathbf{a}}_{n_1 11}). \qquad (16)\end{aligned}$$

Finally, from the definitions of $\mathbf{a}_{13}$, $\mathbf{a}_{12}$, $\mathbf{a}_{23}$, and (10), we derive the relation

$$\mathbf{a}_{13} + \mathbf{a}_{12} + \mathbf{a}_{23} = 2(\overline{\mathbf{a}}_{ijk} - \overline{\mathbf{a}}_{i11} + \overline{\mathbf{a}}_{n_1 11} - \overline{\mathbf{a}}_{n_1 j1} + \overline{\mathbf{a}}_{n_1 n_2 1} - \overline{\mathbf{a}}_{n_1 n_2 k}).$$

If we add (14), (15), and (16), we obtain

$$\begin{aligned}\overline{\mathbf{a}}_{ijk} &= \overline{\mathbf{a}}_{i11} - \overline{\mathbf{a}}_{n_1 11} + \overline{\mathbf{a}}_{n_1 j1} - \overline{\mathbf{a}}_{n_1 n_2 1} + \overline{\mathbf{a}}_{n_1 n_2 k} \\ &\quad - \frac{A_2}{B_2}(\overline{\mathbf{a}}_{n_1 n_2 n_3} - \overline{\mathbf{a}}_{n_1 1 n_3} + \overline{\mathbf{a}}_{n_1 11} - \overline{\mathbf{a}}_{n_1 n_2 1}) \\ &\quad - \frac{A_3}{B_3}(\overline{\mathbf{a}}_{n_1 1 n_3} - \overline{\mathbf{a}}_{11 n_3} + \overline{\mathbf{a}}_{111} - \overline{\mathbf{a}}_{n_1 11}) \\ &\quad - \frac{A_1}{B_1}(\overline{\mathbf{a}}_{111} - \overline{\mathbf{a}}_{1 n_2 1} + \overline{\mathbf{a}}_{n_1 n_2 1} - \overline{\mathbf{a}}_{n_1 11}).\end{aligned}$$

To prove the dual feasibility we need to prove that

$$\begin{aligned}&c_{i11} - c_{n_1 11} + c_{n_1 j1} - c_{n_1 n_2 1} + c_{n_1 n_2 k} \\ &- \frac{A_2}{B_2}(c_{n_1 n_2 n_3} - c_{n_1 1 n_3} + c_{n_1 11} - c_{n_1 n_2 1}) \\ &- \frac{A_3}{B_3}(c_{n_1 1 n_3} - c_{11 n_3} + c_{111} - c_{n_1 11}) \\ &- \frac{A_1}{B_1}(c_{111} - c_{1 n_2 1} + c_{n_1 n_2 1} - c_{n_1 11}) - c_{ijk} \le 0,\end{aligned}$$

or, equivalently,

$$\begin{aligned}&(c_{i11} - c_{n_1 11} + c_{n_1 j1} - c_{ij1}) + (c_{ij1} - c_{n_1 j1} + c_{n_1 jk} - c_{ijk}) \\ &\quad + (c_{n_1 j1} - c_{n_1 n_2 1} + c_{n_1 n_2 k} - c_{n_1 jk}) \\ &\quad \le \frac{A_2}{B_2}(c_{n_1 n_2 n_3} - c_{n_1 1 n_3} + c_{n_1 11} - c_{n_1 n_2 1}) \qquad (17) \\ &\quad + \frac{A_3}{B_3}(c_{n_1 1 n_3} - c_{11 n_3} + c_{111} - c_{n_1 11}) + \frac{A_1}{B_1}(c_{111} - c_{1 n_2 1} + c_{n_1 n_2 1} - c_{n_1 11}).\end{aligned}$$

If we fix one of the i, j, k subscripts in c_{ijk}, then, in view of Theorem 4.1 and the results in Table 3, (8) will hold for the remaining two subscripts. Thus we can obtain the following inequalities:

$$c_{i11} - c_{n_1 11} + c_{n_1 j1} - c_{ij1} \le \frac{A_1}{B_1}(c_{111} - c_{1n_2 1} + c_{n_1 n_2 1} - c_{n_1 11}), \tag{18}$$

$$c_{n_1 j1} - c_{n_1 n_2 1} + c_{n_1 n_2 k} - c_{n_1 jk} \le \frac{A_2}{B_2}(c_{n_1 n_2 n_3} - c_{n_1 1 n_3} + c_{n_1 11} - c_{n_1 n_2 1}), \tag{19}$$

$$c_{ij1} - c_{n_1 j1} + c_{n_1 jk} - c_{ijk} \le \frac{A_3}{B_3}(c_{n_1 j n_3} - c_{1 j n_3} + c_{1j1} - c_{n_1 j1}). \tag{20}$$

Also, from the nonpositivity of the (1,1,1)-order divided difference of $c(y_i, z_j, w_k)$, we obtain

$$\frac{\frac{c_{111} - c_{n_1 11} - c_{11n_3} + c_{n_1 1 n_3}}{(y_1 - y_{n_1})(w_1 - w_{n_3})} - \frac{c_{1j1} - c_{n_1 j1} - c_{1jn_3} + c_{n_1 jn_3}}{(y_1 - y_{n_1})(w_1 - w_{n_3})}}{(z_1 - z_j)} \le 0,$$

where $z_1 - z_j < 0$, $y_1 - y_{n_1} < 0$, $w_1 - w_{n_3} < 0$. This implies

$$c_{1j1} - c_{n_1 j1} - c_{1jn_3} + c_{n_1 jn_3} \le c_{111} - c_{n_1 11} - c_{11n_3} + c_{n_1 1 n_3}. \tag{21}$$

By (20), (21), and the inequality $\frac{A_3}{B_3} > 0$, we have

$$c_{ij1} - c_{n_1 j1} + c_{n_1 jk} - c_{ijk} \le \frac{A_3}{B_3}(c_{111} - c_{n_1 11} - c_{11n_3} + c_{n_1 1 n_3}). \tag{22}$$

If we add (18), (19), and (22), we obtain (17). Thus the basis is dual feasible. □

By a similar proof we can obtain the following theorem.

THEOREM 5.2. *Under the same conditions as in Theorem* 5.1*, except that* $(1,1,1)$*-order divided differences of* $c(y_i, z_j, w_k)$ *are nonnegative, the vectors in* $\overline{S}_1 \cup \{\overline{\mathbf{a}}_{n_1 1 n_3}, \overline{\mathbf{a}}_{1 n_2 1}, \overline{\mathbf{a}}_{1 n_2 n_3}\}$ *form a dual feasible basis of the columns of* $\overline{\mathbf{A}}$*.*

Let

$$\begin{aligned}
\overline{S}_2 &= \{\overline{\mathbf{a}}_{i11}, i = 1, \ldots, n_1, \overline{\mathbf{a}}_{n_1 1k}, k = 2, \ldots, n_3, \overline{\mathbf{a}}_{n_1 jn_3}, j = 2, \ldots, n_2\},\\
\overline{S}_3 &= \{\overline{\mathbf{a}}_{1j1}, j = 1, \ldots, n_2, \overline{\mathbf{a}}_{in_2 1}, i = 2, \ldots, n_1, \overline{\mathbf{a}}_{n_1 n_2 k}, k = 2, \ldots, n_3\},\\
\overline{S}_4 &= \{\overline{\mathbf{a}}_{1j1}, j = 1, \ldots, n_2, \overline{\mathbf{a}}_{1n_2 k}, k = 2, \ldots, n_3, \overline{\mathbf{a}}_{in_2 n_3}, i = 2, \ldots, n_1\},\\
\overline{S}_5 &= \{\overline{\mathbf{a}}_{11k}, k = 1, \ldots, n_3, \overline{\mathbf{a}}_{i1n_3}, i = 2, \ldots, n_1, \overline{\mathbf{a}}_{n_1 jn_3}, j = 2, \ldots, n_2\},\\
\overline{S}_6 &= \{\overline{\mathbf{a}}_{11k}, k = 1, \ldots, n_3, \overline{\mathbf{a}}_{1jn_3}, j = 2, \ldots, n_2, \overline{\mathbf{a}}_{in_2 n_3}, i = 2, \ldots, n_1\},\\
\overline{S'}_1 &= \{\overline{\mathbf{a}}_{i11}, i = 1, \ldots, n_1, \overline{\mathbf{a}}_{1j1}, j = 2, \ldots, n_2, \overline{\mathbf{a}}_{11k}, k = 2, \ldots, n_3\},\\
\overline{S'}_2 &= \{\overline{\mathbf{a}}_{in_2 n_3}, i = 1, \ldots, n_1, \overline{\mathbf{a}}_{n_1 jn_3}, j = 1, \ldots, n_2 - 1,\\
&\qquad \overline{\mathbf{a}}_{n_1 n_2 k}, k = 1, \ldots, n_3 - 1\}.
\end{aligned}$$

The vectors of $\overline{S}_i$, $i = 1, \ldots, 6$, $\overline{S'}_1$, and $\overline{S'}_2$ are illustrated in Figure 3. They are represented by the lattice points of the boldface lines.

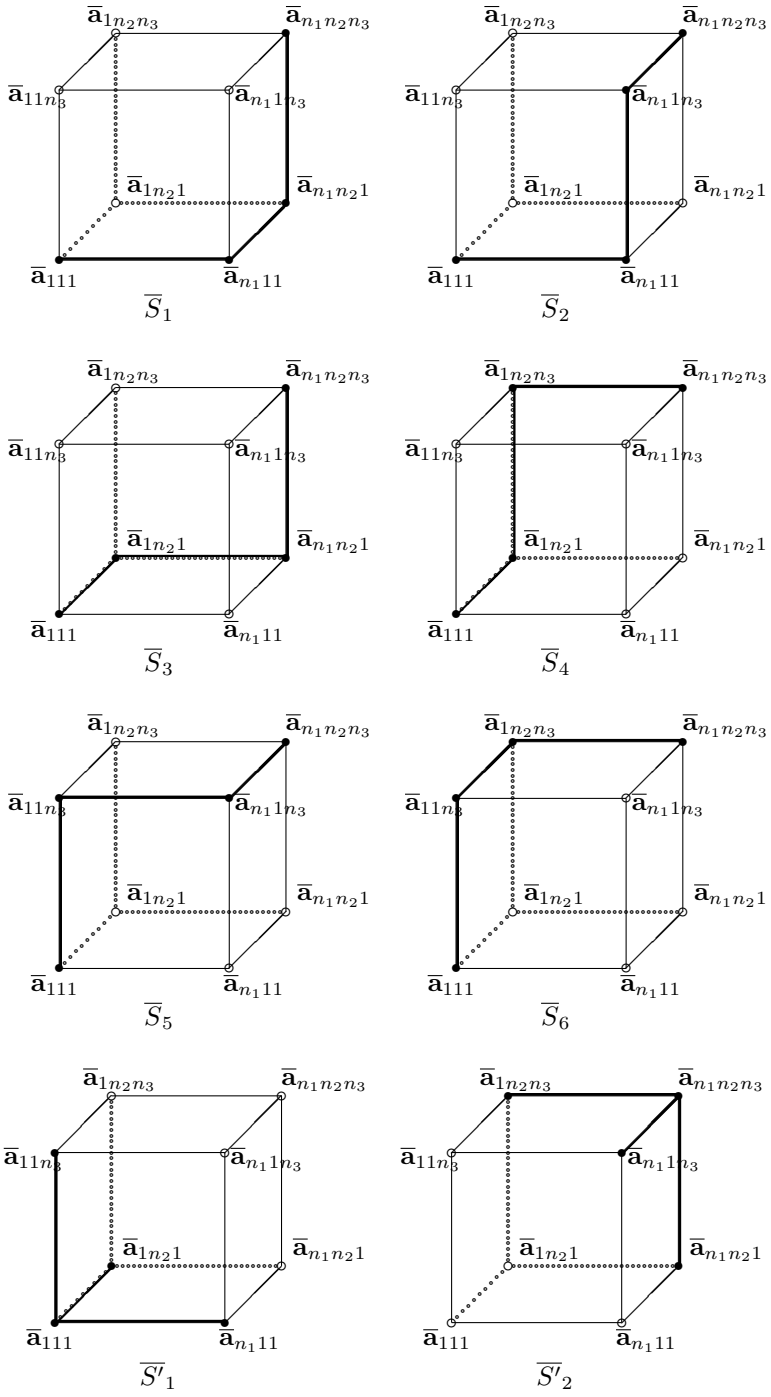

$\overline{\mathbf{a}}_{1n_2n_3}$
$\overline{\mathbf{a}}_{n_1n_2n_3}$
$\overline{\mathbf{a}}_{11n_3}$
$\overline{\mathbf{a}}_{n_11n_3}$
$\overline{\mathbf{a}}_{1n_21}$
$\overline{\mathbf{a}}_{n_1n_21}$
$\overline{\mathbf{a}}_{111}$
$\overline{\mathbf{a}}_{n_111}$
$\overline{S}_1$
$\overline{S}_2$
$\overline{S}_3$
$\overline{S}_4$
$\overline{S}_5$
$\overline{S}_6$
$\overline{S'}_1$
$\overline{S'}_2$

FIG. 3.

TABLE 4
"(i,j,k)" means (i,j,k)-order divided difference of c, $0 \le i,j,k \le 2$. Other notation as in Table 3.

	y_i	↗	↗	↗	↘	↗	↘	↘	↘
	z_j	↗	↗	↘	↗	↘	↗	↘	↘
	w_k	↗	↘	↗	↗	↘	↘	↗	↘
r1	$(1,2,0)$	≥ 0	≥ 0	≥ 0	≤ 0	≥ 0	≤ 0	≤ 0	≤ 0
r2	$(2,1,0)$	≤ 0	≤ 0	≥ 0	≤ 0	≥ 0	≤ 0	≥ 0	≥ 0
r3	$(1,0,2)$	≥ 0	≥ 0	≥ 0	≤ 0	≥ 0	≤ 0	≤ 0	≤ 0
r4	$(2,0,1)$	≤ 0	≥ 0	≤ 0	≤ 0	≥ 0	≥ 0	≤ 0	≥ 0
r5	$(0,1,2)$	≥ 0	≥ 0	≤ 0	≥ 0	≤ 0	≥ 0	≤ 0	≤ 0
r6	$(0,2,1)$	≤ 0	≥ 0	≤ 0	≤ 0	≥ 0	≥ 0	≤ 0	≥ 0
r7	$(1,1,1)$	≤ 0 (≥ 0)				≥ 0 (≤ 0)			
	d.f.b. in min	$\overline{S}_{11}$ $(\overline{S}_{12})$							
	d.f.b. in max	$\overline{S}_{62}$ $(\overline{S}_{61})$							

TABLE 5

Inequalities in Table 4 reversed in rows	d.f.b. in min	d.f.b. in max
r1, r2, r3, r4, r5, r6, r7	$\overline{S}_{62}$ $(\overline{S}_{61})$	$\overline{S}_{11}$ $(\overline{S}_{12})$
r5, r6,	$\overline{S}_{21}$ $(\overline{S}_{22})$	$\overline{S}_{42}$ $(\overline{S}_{41})$
r1, r2, r3, r4, r7	$\overline{S}_{42}$ $(\overline{S}_{41})$	$\overline{S}_{21}$ $(\overline{S}_{22})$
r1, r2,	$\overline{S}_{31}$ $(\overline{S}_{32})$	$\overline{S}_{52}$ $(\overline{S}_{51})$
r3, r4, r5, r6, r7	$\overline{S}_{52}$ $(\overline{S}_{51})$	$\overline{S}_{31}$ $(\overline{S}_{32})$
r1, r3, r5, r7	$\overline{S'}_{11}$ $(\overline{S'}_{12}$, $\overline{S'}_{13}$ or $\overline{S'}_{14})$	$\overline{S'}_{22}$, $\overline{S'}_{23}$ or $\overline{S'}_{24}$ $(\overline{S'}_{21})$
r2, r4, r6	$\overline{S'}_{22}$, $\overline{S'}_{23}$ or $\overline{S'}_{24}$ $(\overline{S'}_{21})$	$\overline{S'}_{11}$ $(\overline{S'}_{12}$, $\overline{S'}_{13}$ or $\overline{S'}_{14}$. $)$

Let

$$\begin{aligned}
\overline{S}_{11} &= \overline{S}_1 \cup \{\overline{\mathbf{a}}_{n_1 1 n_3}, \overline{\mathbf{a}}_{1 n_2 1}, \overline{\mathbf{a}}_{11 n_3}\}, & \overline{S}_{12} &= \overline{S}_1 \cup \{\overline{\mathbf{a}}_{n_1 1 n_3}, \overline{\mathbf{a}}_{1 n_2 1}, \overline{\mathbf{a}}_{1 n_2 n_3}\},\\
\overline{S}_{21} &= \overline{S}_2 \cup \{\overline{\mathbf{a}}_{11 n_3}, \overline{\mathbf{a}}_{n_1 n_2 1}, \overline{\mathbf{a}}_{1 n_2 1}\}, & \overline{S}_{22} &= \overline{S}_2 \cup \{\overline{\mathbf{a}}_{11 n_3}, \overline{\mathbf{a}}_{n_1 n_2 1}, \overline{\mathbf{a}}_{1 n_2 n_3}\},\\
\overline{S}_{31} &= \overline{S}_3 \cup \{\overline{\mathbf{a}}_{n_1 11}, \overline{\mathbf{a}}_{1 n_2 n_3}, \overline{\mathbf{a}}_{11 n_3}\}, & \overline{S}_{32} &= \overline{S}_3 \cup \{\overline{\mathbf{a}}_{n_1 11}, \overline{\mathbf{a}}_{1 n_2 n_3}, \overline{\mathbf{a}}_{n_1 1 n_3}\},\\
\overline{S}_{41} &= \overline{S}_4 \cup \{\overline{\mathbf{a}}_{11 n_3}, \overline{\mathbf{a}}_{n_1 n_2 1}, \overline{\mathbf{a}}_{n_1 11}\}, & \overline{S}_{42} &= \overline{S}_4 \cup \{\overline{\mathbf{a}}_{11 n_3}, \overline{\mathbf{a}}_{n_1 n_2 1}, \overline{\mathbf{a}}_{n_1 1 n_3}\},\\
\overline{S}_{51} &= \overline{S}_5 \cup \{\overline{\mathbf{a}}_{n_1 11}, \overline{\mathbf{a}}_{1 n_2 n_3}, \overline{\mathbf{a}}_{1 n_2 1}\}, & \overline{S}_{52} &= \overline{S}_5 \cup \{\overline{\mathbf{a}}_{n_1 11}, \overline{\mathbf{a}}_{1 n_2 n_3}, \overline{\mathbf{a}}_{n_1 n_2 1}\},\\
\overline{S}_{61} &= \overline{S}_6 \cup \{\overline{\mathbf{a}}_{1 n_2 1}, \overline{\mathbf{a}}_{n_1 1 n_3}, \overline{\mathbf{a}}_{n_1 11}\}, & \overline{S}_{62} &= \overline{S}_6 \cup \{\overline{\mathbf{a}}_{1 n_2 1}, \overline{\mathbf{a}}_{n_1 1 n_3}, \overline{\mathbf{a}}_{n_1 n_2 1}\},\\
\overline{S'}_{21} &= \overline{S'}_2 \cup \{\overline{\mathbf{a}}_{n_1 11}, \overline{\mathbf{a}}_{1 n_2 1}, \overline{\mathbf{a}}_{11 n_3}\}, & \overline{S'}_{22} &= \overline{S'}_2 \cup \{\overline{\mathbf{a}}_{n_1 11}, \overline{\mathbf{a}}_{1 n_2 1}, \overline{\mathbf{a}}_{111}\},\\
\overline{S'}_{23} &= \overline{S'}_2 \cup \{\overline{\mathbf{a}}_{n_1 11}, \overline{\mathbf{a}}_{111}, \overline{\mathbf{a}}_{11 n_3}\}, & \overline{S'}_{24} &= \overline{S'}_2 \cup \{\overline{\mathbf{a}}_{111}, \overline{\mathbf{a}}_{1 n_2 1}, \overline{\mathbf{a}}_{11 n_3}\},
\end{aligned}$$

$$\begin{aligned}
\overline{S'}_{11} &= \overline{S'}_1 \cup \{\overline{\mathbf{a}}_{n_1 n_2 1}, \overline{\mathbf{a}}_{n_1 1 n_3}, \overline{\mathbf{a}}_{1 n_2 n_3}\},\\
\overline{S'}_{12} &= \overline{S'}_1 \cup \{\overline{\mathbf{a}}_{n_1 n_2 1}, \overline{\mathbf{a}}_{n_1 1 n_3}, \overline{\mathbf{a}}_{n_1 n_2 n_3}\},\\
\overline{S'}_{13} &= \overline{S'}_1 \cup \{\overline{\mathbf{a}}_{n_1 n_2 1}, \overline{\mathbf{a}}_{n_1 n_2 n_3}, \overline{\mathbf{a}}_{1 n_2 n_3}\},\\
\overline{S'}_{14} &= \overline{S'}_1 \cup \{\overline{\mathbf{a}}_{n_1 n_2 n_3}, \overline{\mathbf{a}}_{n_1 1 n_3}, \overline{\mathbf{a}}_{1 n_2 n_3}\}.
\end{aligned}$$

These sets can be dual feasible bases of the three-dimensional problem (9) under different conditions. We present some results in Table 4.

There are, however, cases other than those presented in Table 4. We summarize the results for them in Table 5.

6. Applications and illustrative examples. Monge and inverse Monge arrays come up in many practical applications. In this section we present three more applications which, at the same time, illustrate the ways we can use the results of the present paper.

6.1. Bounding unknown entries in partially known arrays. Any dual feasible basis in an LP may serve for bounding and approximation of unknown components of the coefficient vector of the objective function. If $\mathbf{B}_1$ ($\mathbf{B}_2$) is a dual feasible basis in a minimization (maximization) problem such that $\mathbf{c}_{\mathbf{B}_1}$ ($\mathbf{c}_{\mathbf{B}_2}$) is known, then we have the bound for any unknown c_k:

$$\mathbf{y}_{\mathbf{B}_1}^T \mathbf{a}_k \leq c_k (\leq \mathbf{y}_{\mathbf{B}_2}^T \mathbf{a}_k), \tag{23}$$

where $\mathbf{y}_{\mathbf{B}_i}$ is the solution of the equation $\mathbf{y}_{\mathbf{B}_i}^T \mathbf{B}_i = \mathbf{c}_{\mathbf{B}_i}^T$, $i = 1, 2$.

In section 2 we presented dual feasible bases for problem (1) with Monge arrays in the objective function. Each dual feasible basis gives us a lower (upper) bound. Thus, we have created a method for bounding the entries of the above-mentioned arrays if they are only partially known. If both the lower and upper bounds can be given for c_k and the bounds are close, then they may be used for the approximation of that value.

We can improve on the bounds if we take further dual feasible bases, compute the lower (upper) bounds to the unknown c_{ij} values with each of them, and then take the best lower (upper) bounds obtained that way. One way to get further dual feasible bases is to use the $(GREEDY\ DUAL)_d$ algorithm. We make subsequent steps until the algorithm stops either because a primal feasible basis has been found or because the objective function coefficient corresponding to the entering vector is unknown. Then we take the best bound so far as our final (lower or upper) bound.

Any distribution array is a Monge array, and the entries of a distribution array are values of a probability distribution function. Hence, the methodology of sections 2 and 3 provides us with a methodology for bounding and approximation of unknown values of a multivariate discrete probability distribution function.

Example 6.1. We look at problem (1) where the values of partially known Monge array $\mathbf{c}$ and a_i, b_j are given in the Table 6. First we choose the initial dual feasible basis of the minimization problem as

$$L_1 = \{\mathbf{a}_{11}, \mathbf{a}_{12}, \mathbf{a}_{22}, \mathbf{a}_{23}, \mathbf{a}_{33}, \mathbf{a}_{34}, \mathbf{a}_{35}, \mathbf{a}_{36}, \mathbf{a}_{46}\}.$$

The $(GREEDY\ DUAL)_2$ algorithm provides us with the optimal basis in one step by substituting $\mathbf{a}_{24}$ for $\mathbf{a}_{33}$. By the use of the two dual feasible bases encountered in the algorithm, we have obtained the following final lower bounds:

$$c_{13} \geq 11, \quad c_{14} \geq 9, \quad c_{15} \geq 5, \quad c_{16} \geq 8, \quad c_{21} \geq 14, \quad c_{25} \geq 11, \quad c_{26} \geq 14,$$
$$c_{31} \geq 13, \quad c_{32} \geq 15, \quad c_{41} \geq 14, \quad c_{42} \geq 16, \quad c_{43} \geq 17, \quad c_{44} \geq 14, \quad c_{45} \geq 10.$$

TABLE 6
Values used in Example 6.1.

$a_4 = \frac{14}{93}$	c_{41}	c_{42}	c_{43}	c_{44}	c_{45}	$c_{46} = 13$
$a_3 = \frac{39}{93}$	c_{31}	c_{32}	$c_{33} = 16$	$c_{34} = 13$	$c_{35} = 9$	$c_{36} = 12$
$a_2 = \frac{22}{93}$	c_{21}	$c_{22} = 16$	$c_{23} = 17$	$c_{24} = 15$	c_{25}	c_{26}
$a_1 = \frac{18}{93}$	$c_{11} = 8$	$c_{12} = 10$	c_{13}	c_{14}	c_{15}	c_{16}
	$b_1 = \frac{10}{93}$	$b_2 = \frac{11}{93}$	$b_3 = \frac{13}{93}$	$b_4 = \frac{20}{93}$	$b_5 = \frac{24}{93}$	$b_6 = \frac{15}{93}$

TABLE 7
Values of c_{ij} for Example 6.2.

$c_{41} = 0.17$	$c_{42} = 0.34$	$c_{43} = 0.53$	$c_{44} = 0.68$	$c_{45} = 0.81$	$c_{46} = 1.00$
$c_{31} = 0.12$	$c_{32} = 0.23$	$c_{33} = 0.36$	$c_{34} = 0.47$	$c_{35} = 0.57$	$c_{36} = 0.72$
$c_{21} = 0.07$	$c_{22} = 0.14$	$c_{23} = 0.23$	$c_{24} = 0.30$	$c_{25} = 0.37$	$c_{26} = 0.49$
$c_{11} = 0.02$	$c_{12} = 0.05$	$c_{13} = 0.09$	$c_{14} = 0.12$	$c_{15} = 0.16$	$c_{16} = 0.22$

6.2. Bounding a probability $P(\mathbf{X_1} \leq \mathbf{X_2})$. Let $\mathbf{X}_1$, $\mathbf{X}_2$ be two independent random vectors with the same discrete support set. Suppose that the probability distribution of $\mathbf{X}_1$ is fully known but the distribution of $\mathbf{X}_2$ is only partially known. We may know all univariate marginal distributions of the components of $\mathbf{X}_2$ and, if $d = 2$ or 3, all covariances of the pairs. Then we can give lower and upper bounds for $P(\mathbf{X}_1 \leq \mathbf{X}_2)$. Each lower or upper bound is based on a dual feasible basis. Any dual feasible basis that we have presented in sections 2–5 provides us with a bound. If the basis is both primal and dual feasible, then the bound is sharp, and no better bound can be given based on the available information. If only the marginals are known, then the $(GREEDY\ DUAL)_d$ algorithm provides us with the sharp bound. If, however, the covariances are also known, then the dual algorithm can be used to obtain the sharp bounds. The dual feasible bases presented in section 2, 4, 5 can serve as initial bases.

Example 6.2. Suppose that the two-dimensional random vectors $\mathbf{X}_1 = (Y_1, Z_1)$ and $\mathbf{X}_2 = (Y_2, Z_2)$ can take on the values (y_i, z_j), $i = 1, \ldots, 4$, $j = 1, \ldots, 6$, where

$$y_1 = 1, \quad y_2 = 3, \quad y_3 = 4, \quad y_4 = 5,$$
$$z_1 = 18, \quad z_2 = 11, \quad z_3 = 6, \quad z_4 = 3, \quad z_5 = 2, \quad z_6 = 1.$$

The values $c_{ij} = F_{\mathbf{X}_1}(y_i, z_j)$ are given in Table 7.

The distribution of $\mathbf{X}_2$ is not completely known; we know only the univariate marginal distributions $P(Y_2 = y_i) = a_i$, $P(Z_2 = z_j) = b_j$ that are given as follows:

$$a_1 = 0.19, \quad a_2 = 0.24, \quad a_3 = 0.42, \quad a_4 = 0.15,$$
$$b_1 = 0.11, \quad b_2 = 0.12, \quad b_3 = 0.14, \quad b_4 = 0.21, \quad b_5 = 0.26, \quad b_6 = 0.16. \tag{24}$$

Since

$$P(\mathbf{X}_1 \leq \mathbf{X}_2) = E[P(\mathbf{X}_1 \leq \mathbf{X}_2 | \mathbf{X}_2)] = \sum_{i=1}^{4} \sum_{j=1}^{6} c_{ij} F_{\mathbf{X}_2}(y_i, z_j),$$

the lower and upper bounds for $P(\mathbf{X}_1 \leq \mathbf{X}_2)$ can be obtained by application of $(GREEDY\ DUAL)_2$ with the a_i, b_j, and c_{ij} values presented in (24) and Table 7. The inverse Monge property of the array c_{ij} allowed us the choice of an initial dual feasible basis for the minimization problem and one for the maximization problem. The obtained bounds for $P(\mathbf{X}_1 \leq \mathbf{X}_2)$ are as follows:

$$0.3232 \leq P(\mathbf{X}_1 \leq \mathbf{X}_2) \leq 0.4381.$$

We can improve on these bounds if the covariance of Y_2, Z_2 is also known. Assume that this covariance equals -5.003. Since

$$E[Y_2] = \sum_{i=1}^{4} y_i a_i = 3.34, \qquad E[Z_2] = \sum_{j=1}^{6} z_j b_j = 5.45,$$

it follows that

$$E[Y_2Z_2] = \sum_{i=1}^{4}\sum_{j=1}^{6} y_i z_j P_{X_2}(y_i, z_j) = \mathrm{Cov}(Y_2, Z_2) + E[Y_2]E[Z_2]$$
$$= -5.003 + (3.34)(5.45) = 13.2.$$

The new bounds can be obtained by the solutions of the minimization and maximization problems (7) by the use of a dual method with 13.2 on the r.h.s. of the last constraint.

The $c(y_i, z_j) = c_{ij}$ has nonnegative (1,2)-order and nonpositive (2,1)-order divided differences; thus we could choose initial dual feasible bases for those problems by the use of Table 3. Then the application of the dual method for both problems gives us the improved bounds as follows:

$$0.4084 \le P(X_1 \le X_2) \le 0.4337.$$

6.3. The Wasserstein distance of two probability distributions. The Wasserstein distance between the probability distributions μ and ν, defined in R^n, is the value

$$W(\mu, \nu) = \inf_{\pi \in \Pi} \sqrt{\int\int \frac{1}{2} d(x, y)^2 d\pi(x, y)}, \tag{25}$$

where Π is the set of all probability distributions in $R^n \times R^n$ with marginals μ and ν, respectively; i.e., if $\pi \in \Pi$, then $\pi(\circ \times R^n) = \mu$, $\pi(R^n \times \circ) = \nu$, and $d(x, y)$ is the Euclidean distance between x and y.

Let $n = 1$ and μ, ν be the discrete distributions with supports $\{y_1, \ldots, y_m\}$, $\{z_1, \ldots, z_n\}$ and corresponding probabilities $\{a_i\}$, $\{b_j\}$, respectively. Then

$$\begin{aligned} W^2(\mu, \nu) = \quad & \min \quad \textstyle\sum_{i=1}^{m}\sum_{j=1}^{n} \frac{1}{2}(y_i - z_j)^2 x_{ij} \\ & \text{subject to} \\ & \textstyle\sum_{j=1}^{n} x_{ij} = a_i, \quad i = 1, \ldots, m, \\ & \textstyle\sum_{i=1}^{m} x_{ij} = b_j, \quad j = 1, \ldots, n, \\ & x_{ij} \ge 0, \quad i = 1, \ldots, m, \quad j = 1, \ldots, n. \end{aligned} \tag{26}$$

Assume that both $\{y_i\}$ and $\{z_j\}$ are increasing sequences. It is easy to see that the array

$$c_{ij} = \frac{1}{2}(y_i - z_j)^2, \quad i = 1, \ldots, m, \quad j = 1, \ldots, n,$$

has the Monge property (its $(1, 1)$-order divided difference is constant, equal to -1).

By Theorem 2.1 the objective function value, corresponding to any ordered sequence of the minimization transportation problem, is a lower bound on $W^2(\mu, \nu)$. That ordered sequence, which is also primal feasible, provides us with the exact value of $W^2(\mu, \nu)$.

7. Conclusions. We have reformulated some basic results, and obtained them in connection with Monge and inverse Monge arrays, in terms of dual feasible bases in the multidimensional transportation problem. We have also obtained general results in connection with the structures of dual feasible bases in the two- and three-dimensional

cases, where the constraints of the transportation problems are supplemented by covariance constraints.

Our results allow for creating lower and upper bounds for unknown entries in partially known Monge and inverse Monge arrays. We have also shown how the results can be used to obtain lower and upper bounds for the expectation of a function of a discrete random vector, where the values of this function form a Monge or inverse Monge array and the univariate marginal distributions of the random vector are known. In the two- and three-dimensional cases we have obtained improved bounds under the condition that the covariances of the pairs of the random variables are also known. In this case the coefficient array of the objective function is supposed to have some special higher order convexity property.

We have presented three applications of the results of the paper. The first one provides us with bounds for unknown entries in a Monge array. The second one gives bounds for the probability that one random vector dominates another one. In the third one we obtain bounds for the Wasserstein distance between two discrete probability distributions.

REFERENCES

[1] W. W. Bein, P. Brucker, J. K. Park, and P. K. Pathak, *A Monge property for the d-dimensional transportation problem*, Discrete Appl. Math., 58 (1995), pp. 97–109.

[2] R. E. Burkard, *Monge Properties, Discrete Convexity and Applications*, SRC F-003 Technical Report 328, Graz University of Technology, Graz, Austria, 2004.

[3] R. E. Burkard, B. Klinz, and R. Rudolf, *Perspectives of Monge properties in optimization*, Discrete Appl. Math., 70 (1996), pp. 95–161.

[4] A. J. Hoffman, *On simple linear programming problems*, in Proceedings of the Symposium on Pure Mathematics, V. Klee, ed., Convexity VII, 1963, pp. 317–327.

[5] G. Mádi-Nagy and A. Prékopa, *On multivariate discrete moment problems and their applications to bounding expectations and probabilities*, Math. Oper. Res., 29 (2004), pp. 229–258.

[6] A. Prékopa, *A brief introduction to linear programming*, Math. Sci., 21 (1996), pp. 85–111.

[7] A. Prékopa, *Discrete higher order convex functions and their applications*, Generalized Convexity and Generalized Monotonicity, N. Hadjisavvas, J. E. Martinez-Legaz, and J.-P. Penot, eds., Lecture Notes in Econom. and Math. Systems 502, Springer, New York, 2001, pp. 21–47.

SIAM J. OPTIM.
Vol. 18, No. 1, pp. 156–185

SECOND-ORDER CONE PROGRAMMING RELAXATION OF SENSOR NETWORK LOCALIZATION*

PAUL TSENG†

In memory of Jos Sturm

Abstract. The sensor network localization problem has been much studied. Recently Biswas and Ye proposed a semidefinite programming (SDP) relaxation of this problem which has various nice properties and for which a number of solution methods have been proposed. Here, we study a second-order cone programming (SOCP) relaxation of this problem, motivated by its simpler structure and its potential to be solved faster than SDP. We show that the SOCP relaxation, though weaker than the SDP relaxation, has nice properties that make it useful as a problem preprocessor. In particular, sensors that are uniquely positioned among interior solutions of the SOCP relaxation are accurate up to the square root of the distance error. Thus, these sensors, which are easily identified, are accurately positioned. In our numerical simulation, the interior solution found can accurately position up to 80–90% of the sensors. We also propose a smoothing coordinate gradient descent method for finding an interior solution that is faster than an interior-point method.

Key words. sensor network localization, semidefinite program, second-order cone program, approximation algorithm, error bound

AMS subject classifications. 90C22, 90C25, 90C26, 90C27, 90C31, 90C35, 90C59

DOI. 10.1137/050640308

1. Introduction.

A problem that has received considerable attention is that of ad hoc wireless sensor network localization [3, 10, 11, 16, 17, 22, 28, 30, 31]. The basic version of this problem can be described as follows:

> There are n distinct points in $\Re^d$ ($d \geq 1$). We know the Cartesian coordinates of the last $n-m$ points ("anchors") $x_{m+1}, \ldots, x_n$ and the Euclidean distance $d_{ij} > 0$ between "neighboring" points i and j for $(i,j) \in \mathcal{A}$, where $\mathcal{A} \subseteq (\{1,\ldots,n\} \times \{1,\ldots,m\}) \cup (\{1,\ldots,m\} \times \{1,\ldots,n\})$.[1] We wish to estimate the Cartesian coordinates of the first m points ("sensors").

Typically, $d = 2$ and two points are neighbors if the distance between them is below some threshold (the radio range). In variants of this problem, the distances may be non-Euclidean [30] or may have measurement errors, and there may be additional constraints on the unknown points [16]. This problem is closely related to distance geometry problems arising in the determination of protein structure [8, 24] and to graph rigidity [1, 17, 31].

It is known that the sensor network localization problem is NP-hard in general [29]; see also the remark in [24]. A proof for $d = 1$ is by reduction from the set partition problem, which is readily generalized to $d > 1$. Additional studies are given in [3, 28]. Thus, efforts have been directed at solving this problem approximately. A method based on second-order cone programming (SOCP) relaxation was proposed in [16]. In the case where the anchors lie on the "perimeter," a distributed relaxation

*Received by the editors September 14, 2005; accepted for publication (in revised form) September 10, 2006; published electronically February 26, 2007. This research is supported by National Science Foundation grant DMS-0511283.

http://www.siam.org/journals/siopt/18-1/64030.html

†Department of Mathematics, University of Washington, Seattle, WA 98195-4350 (tseng@math.washington.edu).

[1]The set $\mathcal{A}$ is undirected in the sense that $(i,j) = (j,i)$ and $d_{ij} = d_{ji}$ for all $(i,j) \in \mathcal{A}$.

method was proposed in [28]. The performances of these methods were tested through simulations.

Recently, Biswas and Ye proposed an approach to sensor network localization based on semidefinite programming (SDP) relaxation [10, 11]. In this approach, the problem is formulated as the following nonconvex minimization problem:

$$\upsilon_{\text{opt}} \stackrel{\text{def}}{=} \min_{x_1,\ldots,x_m} \sum_{(i,j)\in\mathcal{A}} |\|x_i - x_j\|^2 - d_{ij}^2|, \tag{1}$$

where $\|\cdot\|$ denotes the Euclidean norm. Introduce

$$X \stackrel{\text{def}}{=} [x_1 \ \cdots \ x_m], \quad A \stackrel{\text{def}}{=} [x_{m+1} \ \cdots \ x_n].$$

Then, for each $(i,j) \in \mathcal{A}$,

$$\begin{aligned}
\|x_i - x_j\|^2 &= (x_i - x_j)^T (x_i - x_j) \\
&= (e_i - e_j)^T \begin{bmatrix} X^T \\ A^T \end{bmatrix} \begin{bmatrix} X & A \end{bmatrix} (e_i - e_j) \\
&= b_{ij}^T \begin{bmatrix} X^T \\ I_d \end{bmatrix} \begin{bmatrix} X & I_d \end{bmatrix} b_{ij} \\
&= \left\langle b_{ij} b_{ij}^T, \begin{bmatrix} X^T X & X^T \\ X & I_d \end{bmatrix} \right\rangle_F,
\end{aligned}$$

where e_i is the ith coordinate vector in $\Re^n$ and $b_{ij} \stackrel{\text{def}}{=} \left[\begin{smallmatrix} I_m & 0 \\ 0 & A \end{smallmatrix}\right] (e_i - e_j)$. Throughout, I_k is the $k \times k$ identity matrix and $\langle A, B\rangle_F \stackrel{\text{def}}{=} \text{trace}[AB]$ for any symmetric real matrices A, B of the same dimension. It is not difficult to see that

$$\begin{bmatrix} Y & X^T \\ X & I_d \end{bmatrix} \succeq 0 \text{ has rank } d \quad \Longleftrightarrow \quad Y = X^T X.^2$$

Thus (1) may be reformulated as

$$\begin{aligned}
\upsilon_{\text{opt}} = \ & \min_{Z} \sum_{(i,j)\in\mathcal{A}} \left|\left\langle b_{ij} b_{ij}^T, Z\right\rangle_F - d_{ij}^2\right| \\
& \text{s.t. } [Z_{ij}]_{i,j\geq n-d} = I_d, \ Z \succeq 0, \ \text{rank}\, Z = d.
\end{aligned} \tag{2}$$

Relaxing the rank-d constraint yields the convex problem

$$\begin{aligned}
\upsilon_{\text{sdp}} \stackrel{\text{def}}{=} \ & \min_{Z} \sum_{(i,j)\in\mathcal{A}} \left|\left\langle b_{ij} b_{ij}^T, Z\right\rangle_F - d_{ij}^2\right| \\
& \text{s.t. } [Z_{ij}]_{i,j\geq n-d} = I_d, \ Z \succeq 0,
\end{aligned} \tag{3}$$

which is an SDP. In particular, by introducing slack variables, this can be written in the standard conic form

$$\begin{aligned}
\min \quad & \sum_{(i,j)\in\mathcal{A}} u_{ij} + v_{ij} \\
\text{s.t.} \quad & \left\langle b_{ij} b_{ij}^T, Z\right\rangle_F - u_{ij} + v_{ij} = d_{ij}^2 \quad \forall (i,j) \in \mathcal{A}, \\
& [Z_{ij}]_{i,j\geq n-d} = I_d, \\
& u_{ij} \geq 0, \ v_{ij} \geq 0 \quad \forall (i,j) \in \mathcal{A}, \ Z \succeq 0,
\end{aligned} \tag{4}$$

[2] In general, $\left[\begin{smallmatrix} u \\ v \end{smallmatrix}\right] \in \text{Null}\left[\begin{smallmatrix} Y & X^T \\ X & I_d \end{smallmatrix}\right]$ if and only if $u \in \text{Null}(Y - X^T X)$, $v = -Xu$.

which has $(m+d)(m+d+1)/2+2|\mathcal{A}|$ variables and $|\mathcal{A}|+d(d+1)/2$ equality constraints. Here $|\mathcal{A}|$ denotes the cardinality of $\mathcal{A}$. In sensor network localization, $|\mathcal{A}| = \Omega(m)$ and $d = 2$ so that (4) has $\Omega(m^2)$ variables and $\Omega(m)$ equality constraints. Properties of the SDP relaxation and its solutions are studied in [10, 31].[3] As noted in [11], the SDP relaxation can be solved by existing SDP solvers for $m \leq 100$ but not for much larger m. Thus, a distributed (domain decomposition) method is proposed to solve larger SDP relaxations. In [22], to further improve the speed and accuracy, the distributed SDP method is terminated early and then a gradient search method is used to locally refine the approximate solution. Simulation results show that this method can more quickly and accurately position most sensors, even in the presence of distance errors.

The challenge in solving the SDP relaxation motivates us to consider SOCP relaxation, first studied by Doherty, Pister, and El Ghaoui [16], since SOCP can be solved to a much larger size than SDP [2, 27]. In fact, there has been little study of SOCP relaxation, compared to SDP relaxation, for nonconvex optimization. Besides [16], which presented models and simulation results with SOCP relaxations of sensor network localization (assuming no distance error), Kim and Kojima [20] and Kim, Kojima, and Yamashita [21] studied SOCP relaxations of certain special classes of SDP and quadratic optimization problems, but their results do not apply to sensor network localization. Here, we present a study, both theoretical and numerical, of the SOCP relaxation of the sensor network localization problem (1), allowing for distance errors. In particular, we show that an interior solution of the SOCP relaxation can be used to accurately position a high percentage of the sensors.[4] To motivate the SOCP relaxation, we reformulate (1) as

$$\begin{aligned} v_{\rm opt} = \quad & \min_{x_1,\ldots,x_m,y_{ij}} \sum_{(i,j)\in\mathcal{A}} |y_{ij} - d_{ij}^2| \\ & \text{s.t.} \;\; y_{ij} = \|x_i - x_j\|^2 \;\; \forall (i,j) \in \mathcal{A}. \end{aligned} \tag{5}$$

Relaxing the equality constraints to "$\geq$" inequality constraints yields the convex problem

$$\begin{aligned} v_{\rm socp} \stackrel{\rm def}{=} \quad & \min_{x_1,\ldots,x_m,y_{ij}} \sum_{(i,j)\in\mathcal{A}} |y_{ij} - d_{ij}^2| \\ & \text{s.t.} \;\; y_{ij} \geq \|x_i - x_j\|^2 \quad \forall (i,j) \in \mathcal{A}, \end{aligned} \tag{6}$$

which is an SOCP. In particular, by noting that $y_{ij} \geq d_{ij}^2$ in any solution of (6) and introducing slack variables, this can be written in the standard conic form

$$\begin{aligned} \min \quad & \sum_{(i,j)\in\mathcal{A}} u_{ij} \\ \text{s.t.} \quad & x_i - x_j - w_{ij} = 0 \quad \forall (i,j) \in \mathcal{A}, \\ & y_{ij} - u_{ij} = d_{ij}^2 \quad \forall (i,j) \in Z\mathcal{A}, \\ & \alpha_{ij} = \tfrac{1}{2} \quad \forall (i,j) \in \mathcal{A}, \\ & u_{ij} \geq 0, \; (\alpha_{ij}, y_{ij}, w_{ij}) \in \text{Rcone}^{d+2} \quad \forall (i,j) \in \mathcal{A}, \end{aligned} \tag{7}$$

where $\text{Rcone}^{d+2} \stackrel{\rm def}{=} \{(\alpha, y, w) \in \Re \times \Re \times \Re^d : 2\alpha y \geq \|w\|^2\}$ [32]. This is an SOCP since

[3] Throughout, "solution" of an optimization problem means a global optimal solution.

[4] Throughout, "interior solution" means an element in the relative interior of the optimal solution set.

$$y \geq \|w\|^2 \iff \left(y+\frac{1}{4}\right)^2 \geq \left(y-\frac{1}{4}\right)^2 + \|w\|^2 \iff y+\frac{1}{4} \geq \left\|\left(y-\frac{1}{4}, w\right)\right\|$$

(see [4, p. 88] or [25, p. 221]). The SOCP (7) has $(d+3)|\mathcal{A}|+md$ variables and $(d+2)|\mathcal{A}|$ equality constraints. In sensor network localization, $|\mathcal{A}| = \Omega(m)$ and $d = 2$, so that (7) has $\Omega(m)$ variables and $\Omega(m)$ equality constraints. Thus, the SOCP relaxation has smaller size than the SDP relaxation.

How good an approximation is the SOCP relaxation? Can it be efficiently solved? We will show that the SOCP relaxation is always weaker than the SDP relaxation and that any interior solution of the SOCP relaxation (which can be found by, say, an interior-point method) will accurately position (up to square root distance error) those sensors that are uniquely positioned; see Propositions 3.1, 7.1, and 7.2. Moreover, the aforementioned sensors (which lie in the convex hull of the anchors) can be easily identified; see Propositions 5.1 and 6.2. In our simulations, described in section 9, up to 80–90% of the sensors are accurately positioned using this technique. Thus, the SOCP relaxation can act as a useful preprocessor by accurately positioning most of the sensors, thus greatly reducing the problem size. The remaining sensors can be positioned by other means, such as SDP relaxation. In section 8, we propose a smoothing coordinate gradient descent (SCGD) method that computes an interior solution of the SOCP relaxation faster than an interior-point method. In sections 10 and 11, we present a mixed SDP-SOCP relaxation of (1), which can flexibly mediate between strength of relaxation and problem size, and discuss alternative problem formulations. In particular, other objective functions can be used in (1), for which SOCP relaxation may be more "natural" than SDP relaxation. However, changing the objective function of (1) changes its solution. Here we consider (1) so to better compare with the existing SDP relaxation approach (Propositions 3.1 and 4.1) and to introduce the mixed SDP-SOCP relaxation. In addition, the SOCP relaxation is a useful problem preprocessor even if it is weaker than SDP relaxation.

Throughout, $\Re^n$ denotes the space of n-dimensional real column vectors (sometimes written horizontally for convenience), $\mathcal{S}^n$ denotes the space of $n \times n$ real symmetric matrices, and ^{T} denotes transpose. For $A \in \Re^{m\times n}$, A_{ij} denotes the (i,j)th entry of A. For $A, B \in \mathcal{S}^n$, $A \succeq B$ means $A - B$ is positive semidefinite. "conv" means the convex hull.

2. An illustrative example. To understand properties of SDP and SOCP relaxations, it is instructive to look at an example. Consider the following example of Ye, with $d = 2$, $n = 3$, $m = 1$, and

$$x_2 = (-1, 0), \quad x_3 = (1, 0), \quad d_{12} = d_{13} = 2.$$

The optimization problem (1) is

$$\min_{x_1=(\alpha,\beta)\in\Re^2} \quad |(1-\alpha)^2 + \beta^2 - 4| + |(-1-\alpha)^2 + \beta^2 - 4|.$$

It has two solutions at $x_1 = (0, \sqrt{3})$, $x_1 = (0, -\sqrt{3})$; see Figure 1.

The SDP relaxation (3) is

$$\begin{array}{ll} \min\limits_{\substack{x_1=(\alpha,\beta)\in\Re^2\\ y\in\Re}} & |y - 2\alpha - 3| + |y + 2\alpha - 3| \\ \text{s.t.} & \begin{bmatrix} y & \alpha & \beta \\ \alpha & 1 & 0 \\ \beta & 0 & 1 \end{bmatrix} \succeq 0. \end{array}$$

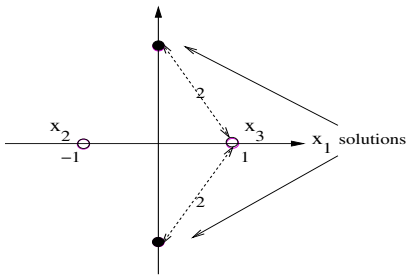


FIG. 1. *The localization problem has two solutions at* $(0, \pm\sqrt{3})$.

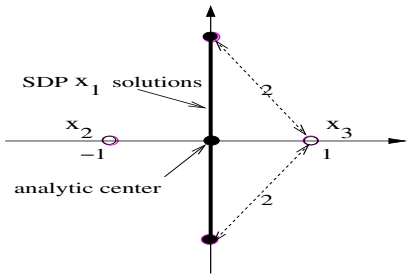


FIG. 2. *The SDP relaxation has the entire line segment as its* x_1*-solution set.*

Its solutions have the form $y = 3$ and x_1 is any point on the line segment joining $(0, -\sqrt{3})$ and $(0, \sqrt{3})$. If we solve the corresponding SDP (4) by an interior-point method, then it will find the solution that maximizes the barrier (see [25, p. 235], [4, p. 384])

$$\det \begin{bmatrix} 3 & 0 & \beta \\ 0 & 1 & 0 \\ \beta & 0 & 1 \end{bmatrix} = 3 - \beta^2.$$

The maximum is attained at $\beta = 0$. The corresponding x_1-solution $(0, 0)$ is the analytic center of the SDP solution set; see Figure 2.

The SOCP relaxation (6) is

$$\begin{array}{ll} \min\limits_{\substack{x_1=(\alpha,\beta)\in\Re^2 \\ y,z\in\Re}} & |y-4| + |z-4| \\ \text{s.t.} & y \geq (\alpha-1)^2 + \beta^2, \\ & z \geq (\alpha+1)^2 + \beta^2. \end{array}$$

Its solutions have the form $y = z = 4$ and x_1 is any point in the intersection of the two disks of radius 2 and centered at $(-1, 0)$ and $(1, 0)$. If we solve the corresponding SOCP (7) by an interior-point method, then it will find the solution that maximizes the barrier (see [25, p. 223], [4, p. 384], and also section 6)

$$\log(4 - (\alpha-1)^2 - \beta^2) + \log(4 - (\alpha+1)^2 - \beta^2).$$

This maximization is attained at $\alpha = \beta = 0$. The corresponding x_1-solution $(0, 0)$ is the analytic center of the SOCP solution set; see Figure 3. In general, finding the

analytic center may be more efficient and accurate than the bounding approach suggested in [16], which entails solving an SOCP $2d$ times with different linear objective functions.

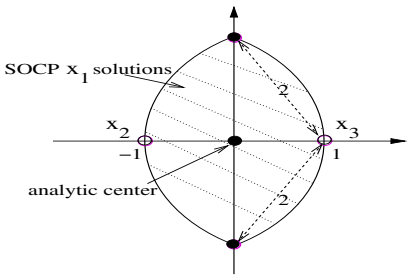


FIG. 3. *The SOCP relaxation has the intersection of two disks as its x_1-solution set.*

From the above example, we make the following observations:

- The SDP x_1-solution set is contained in the SOCP x_1-solution set.
- The analytic center of the SOCP x_1-solution set lies in the convex hull of its neighbors x_2 and x_3.

We will now study in more generality these observed properties of the SDP and SOCP relaxations.

3. Properties of SDP and SOCP relaxations. We show below that the SDP $(x_1, \ldots, x_m)$-solution set is contained in the SOCP $(x_1, \ldots, x_m)$-solution set, so that the SOCP relaxation is weaker than the SDP relaxation.

PROPOSITION 3.1. *If $Z = \left[\begin{smallmatrix} Y & X^T \\ X & I_d \end{smallmatrix}\right]$ is feasible for the SDP relaxation* (3)*, then*

$$x_i = i\text{th column of } X, \quad i = 1, \ldots, m,$$
$$y_{ij} = \begin{cases} Y_{ii} - 2Y_{ij} + Y_{jj} & \text{if } (i,j) \in \mathcal{A},\ i < j \le m; \\ \|x_i\|^2 - 2x_i^T x_j + Y_{jj} & \text{if } (i,j) \in \mathcal{A},\ j \le m < i, \end{cases}$$

is feasible for the SOCP relaxation (6) *with the same objective function value.*

Proof. Since Z is feasible for (3), we have $Z \succeq 0$, so that $Y - X^T X \succeq 0$. Then any 2×2 principal submatrix of $Y - X^T X$ is positive semidefinite, so that, for any $(i,j) \in \mathcal{A}$ with $i < j \le m$,

$$\begin{bmatrix} Y_{ii} - \|x_i\|^2 & Y_{ij} - x_i^T x_j \\ Y_{ij} - x_i^T x_j & Y_{jj} - \|x_j\|^2 \end{bmatrix} \succeq 0,$$

implying that

$$Y_{ii} \ge \|x_i\|^2, \quad Y_{jj} \ge \|x_j\|^2, \quad (Y_{ii} - \|x_i\|^2)(Y_{jj} - \|x_j\|^2) \ge (Y_{ij} - x_i^T x_j)^2. \tag{8}$$

For any $a \ge 0$, $b \ge 0$, $ab \ge c^2$, we have $(a+b)^2 = 4ab + (a-b)^2 \ge 4c^2$ and hence $a + b \ge 2|c|$. Thus (8) implies

$$Y_{ii} - \|x_i\|^2 + Y_{jj} - \|x_j\|^2 \ge 2|Y_{ij} - x_i^T x_j| \ge 2(Y_{ij} - x_i^T x_j).$$

Hence

$$y_{ij} = Y_{ii} - 2Y_{ij} + Y_{jj} \ge \|x_i\|^2 - 2x_i^T x_j + \|x_j\|^2 = \|x_i - x_j\|^2.$$

Similarly, any diagonal entry of $Y - X^T X$ is nonnegative, so that, for any $(i,j) \in \mathcal{A}$ with $j \le m < i$, $Y_{jj} - \|x_j\|^2 \ge 0$ and hence

$$y_{ij} = \|x_i\|^2 - 2x_i^T x_j + Y_{jj} \ge \|x_i\|^2 - 2x_i^T x_j + \|x_j\|^2 = \|x_i - x_j\|^2.$$

Thus $x_1, \ldots, x_m$, $(y_{ij})_{(i,j)\in\mathcal{A}}$ is feasible for (6).

Last, we have from the definition of b_{ij} and y_{ij} that

$$\begin{aligned}
&\sum_{(i,j)\in\mathcal{A}} \left|\left\langle b_{ij} b_{ij}^T, Z\right\rangle_F - d_{ij}^2\right| \\
&= \sum_{\substack{i<j\le m \\ (i,j)\in\mathcal{A}}} |Y_{ii} - 2Y_{ij} + Y_{jj} - d_{ij}^2| + \sum_{\substack{j\le m<i \\ (i,j)\in\mathcal{A}}} |\|x_i\|^2 - 2x_i^T x_j + Y_{jj} - d_{ij}^2| \\
&= \sum_{\substack{i<j\le m \\ (i,j)\in\mathcal{A}}} |y_{ij} - d_{ij}^2| + \sum_{\substack{j\le m<i \\ (i,j)\in\mathcal{A}}} |y_{ij} - d_{ij}^2|.
\end{aligned}$$

Thus, Z and $x_1, \ldots, x_m$, $(y_{ij})_{(i,j)\in\mathcal{A}}$ have the same objective function value for (3) and (6), respectively. □

Proposition 3.1 shows that (i) $v_{\text{SDP}} \ge v_{\text{SOCP}}$ and (ii) if $v_{\text{SDP}} = v_{\text{SOCP}}$, then the set of SDP solutions is contained in the set of SOCP solutions when projected onto the $x_1, \ldots, x_m$-space.

It is well known that the solution set of (3) is closed and convex, and the same is true of (6). An interior solution can be found by, say, applying an interior-point method to (4) and (7). We will see that such an interior solution has desirable properties for identifying sensors that are accurately positioned.

When solving SDP or SOCP by an interior-point method, the solution set must be bounded. It is readily seen that the solution set of (3) is bounded if and only if the solution set of (4) is bounded. Similarly, the solution set of (6) is bounded if and only if the solution set of (7) is bounded. In [31, Proposition 1], it is shown in the case of $v_{\text{opt}} = 0$ (i.e., no distance error) that the solution set of (3) is bounded if the following assumption holds.

ASSUMPTION 1. *Each connected component of the graph* $\mathcal{G} \stackrel{\text{def}}{=} (\{1, \ldots, n\}, \mathcal{A})$ *contains an anchor index.*

It is not difficult to see that this remains true when $v_{\text{opt}} > 0$ and that the converse also holds. Similarly, it is readily shown that the set of solutions of (6) is bounded if and only if Assumption 1 holds. This is summarized in the following lemma.

LEMMA 3.2. (a) *The solution set of* (3) *is bounded if and only if Assumption* 1 *holds.*

(b) *The solution set of* (6) *is bounded if and only if Assumption* 1 *holds.*

Assumption 1 is reasonable since if a connected component of $\mathcal{G}$ does not contain an anchor index, then the corresponding sensors cannot be accurately positioned. In the absence of an anchor (i.e., $m = n$), as arises in protein structure prediction, the solution set is unbounded and, in particular, each solution can be rotated and translated to yield another solution. In [8], an optimization formulation is proposed to remove the translation factor and ensure a bounded solution set (assuming no distance error) and an extension of the distributed SDP method in [11] is proposed, in which points in overlapping "subconfigurations" are further rotated and translated to match closely on the overlap.

4. Interior solution of the SDP relaxation. Let $Z = \left[\begin{smallmatrix} Y & X^T \\ X & I_d \end{smallmatrix}\right]$ be any solution of the SDP relaxation (3). Biswas and Ye introduced the notion of individual traces of Z, defined by

$$\mathrm{tr}_i[Z] \stackrel{\mathrm{def}}{=} Y_{ii} - \|x_i\|^2, \quad i = 1, \ldots, m,$$

where x_i is the ith column of X. Since $Z \succeq 0$ so that $Y - X^T X \succeq 0$, we have

$$\mathrm{tr}_i[Z] \geq 0, \quad i = 1, \ldots, m.$$

These individual traces were given a probabilistic interpretation in [10, section 4] as the variance of random points $\tilde{x}_i$ with $\mathrm{E}[\tilde{x}_i] = x_i$ and $\mathrm{E}[\tilde{x}_i^T \tilde{x}_j] = Y_{ij}$. In [11, section 2], they were used to evaluate the accuracy of the estimated positions x_i, $i = 1, \ldots, m$, with smaller trace indicating higher accuracy. So and Ye [31, Theorem 2] proved in the case of $\upsilon_{\mathrm{sdp}} = 0$ that the sensors are "uniquely localizable" if and only if, for any interior solution Z (equivalently, Z is a solution of maximum rank), all individual traces of Z are zero, i.e., $Y = X^T X$.

The following proposition provides some justification for using individual traces to evaluate accuracy of computed sensor positions. It shows that, for any interior solution of (3), if the ith trace is zero, then the ith sensor is uniquely positioned by the SDP relaxation (and hence is correctly positioned when $\upsilon_{\mathrm{sdp}} = 0$). This result gives a local generalization of the "if" direction in [31, Theorem 2] that is analogous to [31, Theorem 4].

PROPOSITION 4.1. *Let $Z = \left[\begin{smallmatrix} Y & X^T \\ X & I_d \end{smallmatrix}\right]$ be an interior solution of* (3)*. For each $i \in \{1, \ldots, m\}$, if $\mathrm{tr}_i[Z] = 0$, then x_i is invariant over all solutions of* (3)*, where x_i is the ith column of X. Moreover, $Y_{JJ} - X_J^T X_J = 0$, where $J \stackrel{\mathrm{def}}{=} \{i \leq m : \mathrm{tr}_i[Z] = 0\}$, Y_{JJ} is the principal submatrix of Y indexed by J, and X_J is submatrix of X comprising the columns indexed by J.*

Proof. Consider any solution Z' of (3). Since Z is an interior solution, then

$$Z^1 \stackrel{\mathrm{def}}{=} Z + \epsilon(Z' - Z), \quad Z^2 \stackrel{\mathrm{def}}{=} Z - \epsilon(Z' - Z)$$

are both solutions of (3) for any sufficiently small $\epsilon > 0$. Write them in the forms

$$Z^1 = \begin{bmatrix} Y^1 & (X^1)^T \\ X^1 & I_d \end{bmatrix} \quad \text{and} \quad Z^2 = \begin{bmatrix} Y^2 & (X^2)^T \\ X^2 & I_d \end{bmatrix}.$$

Since $Z = (Z^1 + Z^2)/2$, this implies that, for any $i \in \{1, \ldots, m\}$,

$$\begin{aligned} \mathrm{tr}_i[Z] &= \mathrm{tr}_i[(Z^1 + Z^2)/2] \\ &= (Y_{ii}^1 + Y_{ii}^2)/2 - \|(x_i^1 + x_i^2)/2\|^2 \\ &= (Y_{ii}^1 + Y_{ii}^2)/2 - \big(\|x_i^1\|^2 + \|x_i^2\|^2 - \|x_i^1 - x_i^2\|^2/2\big)/2 \\ &= (\mathrm{tr}_i[Z^1] + \mathrm{tr}_i[Z^2])/2 + \|x_i^1 - x_i^2\|^2/4, \end{aligned}$$

where x_i^1, x_i^2 are the ith columns of X^1, X^2, respectively. Since $\mathrm{tr}_i[Z^1] \geq 0$ and $\mathrm{tr}_i[Z^2] \geq 0$, if $\mathrm{tr}_i[Z] = 0$, then $x_i^1 = x_i^2$ and hence $x_i' = x_i$.

Since $Y - X^T X \succeq 0$, we have $Y_{JJ} - X_J^T X_J \succeq 0$. Since $0 = \mathrm{tr}_i[Z] = [Y - X^T X]_{ii}$ for all $i \in J$ so that $Y_{JJ} - X_J^T X_J$ has zero diagonals, this implies $Y_{JJ} - X_J^T X_J = 0$. ∎

Proposition 4.1 shows that any interior solution identifies some subset of sensors that are uniquely positioned by the SDP relaxation. It is an open question whether

the converse of Proposition 4.1 holds, i.e., if Z is an interior solution of (3) and $\mathrm{tr}_i[Z] > 0$, then x_i is not invariant over all solutions of (3). We will prove in section 5 an analogous result for the SOCP relaxation (6).

When an interior-point method is used to solve the SDP relaxation (4), it will find not only an interior solution, but an interior solution that maximizes the nonzero traces in some sense. Using such a solution should make the zero-trace test more robust under computation errors. A rigorous study of this topic requires knowledge of the asymptotic behavior of the central path for SDP, which is not fully understood; see [26] and references therein. On the other hand, the simpler structure of the SOCP relaxation (7) makes possible such a study, as we will do in section 5.

5. Interior solution of the SOCP relaxation. Since the SOCP is a convex minimization problem, there exists a maximal subset of constraints that are tight/active at every solution. In particular, there exists a unique $\mathcal{B} \subseteq \mathcal{A}$ such that

$$\|x_i - x_j\|^2 = y_{ij} \ \forall \text{ solutions } x_1, \ldots, x_m, (y_{ij})_{(i,j)\in\mathcal{A}} \text{ of (6)} \quad \Longleftrightarrow \quad (i,j) \in \mathcal{B}. \tag{9}$$

Any solution that satisfies strictly the remaining constraints of (6) lies in the relative interior of the solution set; i.e., it is an interior solution.

In what follows, we denote the set of neighbors of $i \in \{1, \ldots, m\}$ relative to any $\mathcal{B} \subseteq \mathcal{A}$ by

$$N_{\mathcal{B}}(i) \stackrel{\text{def}}{=} \{j \in \{1, \ldots, n\} \ : \ (i,j) \in \mathcal{B}\}.$$

Also,

$$M_{\mathcal{B}} \stackrel{\text{def}}{=} \{i \in \{1, \ldots, m\} \ : \ N_{\mathcal{B}}(i) \neq \emptyset\}.$$

The next result is key for identifying those sensors that are uniquely positioned by the SOCP relaxation.

PROPOSITION 5.1. *Let $x_1, \ldots, x_m$, $(y_{ij})_{(i,j)\in\mathcal{A}}$ be any interior solution of* (6). *Let $\mathcal{B}$ be given by* (9). *The following results hold.*

(a) *For each $i \in M_{\mathcal{B}}$,*

$$x_i \in \mathrm{conv}\,\{x_j\}_{j\in N_{\mathcal{B}}(i)}\,. \tag{10}$$

(b) *Each connected component of the graph $G_{\mathcal{B}} \stackrel{\text{def}}{=} (M_{\mathcal{B}} \cup \{m+1, \ldots, n\}, \mathcal{B})$ contains an anchor index $i \in \{m+1, \ldots, n\}$.*

(c) *For each $i \in \{1, \ldots, m\}$, x_i is invariant over all solutions of* (6) *if and only if $i \in M_{\mathcal{B}}$.*

Proof. (a) We argue by contradiction. Suppose that (10) fails to hold for some $i \in M_{\mathcal{B}}$. Let p_i denote the nearest-point projection of x_i onto $\mathrm{conv}\,\{x_j\}_{j\in N_{\mathcal{B}}(i)}$. Then, $p_i \neq x_i$ and, for each $j \in N_{\mathcal{B}}(i)$, we have $(x_i - p_i)^T(p_i - x_j) \geq 0$, implying

$$\begin{aligned}\|x_i - x_j\|^2 &= \|x_i - p_i + p_i - x_j\|^2 \\ &= \|x_i - p_i\|^2 + \|p_i - x_j\|^2 + 2(x_i - p_i)^T(p_i - x_j) \\ &> \|p_i - x_j\|^2.\end{aligned}$$

For $\epsilon \in (0,1)$, let

$$x_i^\epsilon = (1-\epsilon)x_i + \epsilon p_i.$$

Since $\|x_i - x_j\|^2 = y_{ij}$ for all $j \in N_{\mathcal{B}}(i)$ and $\|x_i - x_j\|^2 < y_{ij}$ for all $j \in N_{\mathcal{A}\setminus\mathcal{B}}(i)$, the convexity and continuity of $\|\cdot\|^2$ yield that

$$\|x_i^\epsilon - x_j\|^2 < y_{ij} \quad \forall j \in N_{\mathcal{A}}(i),$$

for all ϵ sufficiently small. Thus, replacing x_i by x_i^ϵ yields another solution of (6), and it satisfies strictly the constraints corresponding to $j \in N_{\mathcal{B}}(i)$. This contradicts the assumption that $x_1, \ldots, x_m$, $(y_{ij})_{(i,j)\in\mathcal{A}}$ is an interior solution.

(b) Choose any $\bar{i} \in M_{\mathcal{B}}$ and initialize $\bar{M} \leftarrow \{\bar{i}\}$. Then, whenever there is an $i \in \bar{M} \cap M_{\mathcal{B}}$ with $N_{\mathcal{B}}(i) \not\subseteq \bar{M}$, we add $N_{\mathcal{B}}(i)$ to $\bar{M}$, i.e., $\bar{M} \leftarrow \bar{M} \cup N_{\mathcal{B}}(i)$, until no such i exists. Since, for each $i \in M_{\mathcal{B}}$, each $j \in N_{\mathcal{B}}(i)$ either indexes an anchor or else belongs to $M_{\mathcal{B}}$ (since $N_{\mathcal{B}}(j) \neq \emptyset$), we see that $\bar{M} \subseteq M_{\mathcal{B}} \cup \{m+1, \ldots, n\}$. Moreover, for each $i \in \bar{M} \cap M_{\mathcal{B}}$, we have $N_{\mathcal{B}}(i) \subseteq \bar{M}$ and, by (a), (10) holds, so that

$$x_i \in \operatorname{conv}\{x_j\}_{j\in N_{\mathcal{B}}(i)} \subseteq \operatorname{conv}\{x_j\}_{j\in\bar{M}},$$

implying that x_i is not an extreme point of $\{x_j\}_{j\in\bar{M}}$. Thus, all extreme points of $\operatorname{conv}\{x_j\}_{j\in\bar{M}}$ are anchors. Let

$$\bar{\mathcal{A}} = \{(i,j) \ : \ i \in \bar{M} \cap M_{\mathcal{B}},\ j \in N_{\mathcal{B}}(i)\}.$$

Then $(\bar{M}, \bar{\mathcal{A}})$ is a connected subgraph of $G_{\mathcal{B}}$, and it contains an anchor index; see Figure 4 for an illustrative example. Thus the connected component of $G_{\mathcal{B}}$ that contains this subgraph contains an anchor index. Since the choice of $\bar{i} \in M_{\mathcal{B}}$ was arbitrary, this shows that every connected component of $G_{\mathcal{B}}$ contains an anchor index.

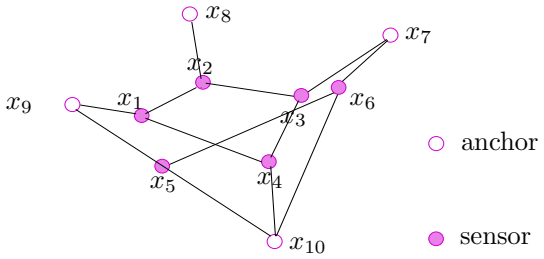


FIG. 4. *In this example, $\mathcal{B}$ is shown as lines and $M_{\mathcal{B}} = \{1, 2, \ldots, 6\}$. For $\bar{i} \in \{1,2,3,4\}$, we have $\bar{M} = \{1,2,3,4,7,8,9,10\}$, $\bar{\mathcal{A}} = \{(1,2),(1,4),(1,9),(2,3),(2,8),(3,7),(3,4),(4,10)\}$. For $\bar{i} \in \{5,6\}$, we have $\bar{M} = \{5,6,7,9,10\}$, $\bar{\mathcal{A}} = \{(5,6),(5,9),(5,10),(6,7),(6,10)\}$.*

(c) If $x'_1, \ldots, x'_m$, $(y'_{ij})_{(i,j)\in\mathcal{A}}$ is any solution of (6), then for each $(i,j) \in \mathcal{B}$,

$$\|x'_i - x'_j\|^2 = y'_{ij}$$

(with $x'_i = x_i$ for $i > m$). Combining this with $\|x_i - x_j\|^2 = y_{ij}$ yields

$$\frac{y_{ij} + y'_{ij}}{2} = \left\| \frac{x_i - x_j}{2} + \frac{x'_i - x'_j}{2} \right\|^2 + \left\| \frac{x_i - x_j}{2} - \frac{x'_i - x'_j}{2} \right\|^2. \tag{11}$$

Since the solution set of (6) is convex, so that $\frac{1}{2}(x_1 + x'_1), \ldots, \frac{1}{2}(x_m + x'_m)$, $(\frac{1}{2}(y_{ij} + y'_{ij}))_{(i,j)\in\mathcal{A}}$ also forms a solution, $(i,j) \in \mathcal{B}$ implies that the rightmost term in (11) must be zero. This in turn implies that

$$x'_i - x'_j = x_i - x_j.$$

Thus, for each $(i,j) \in \mathcal{B}$ there exists $\Delta_{ij} \in \Re^d$ such that

$$x'_i - x'_j = \Delta_{ij} \quad \begin{array}{l} \forall \text{ solutions } x'_1, \ldots, x'_m, (y'_{ij})_{(i,j)\in\mathcal{A}} \\ \text{of (6) (with } x'_i = x_i \text{ for } i > m). \end{array} \tag{12}$$

Let $(\bar{M}, \bar{\mathcal{B}})$ be any connected component of $G_{\mathcal{B}}$. By (b), there exists $i \in \bar{M} \cap \{m+1, \ldots, n\}$, i.e., x_i is an anchor. For each $j \in N_{\bar{\mathcal{B}}}(i)$, we have $(i,j) \in \mathcal{B}$ so that (12) implies

$$x_i - x'_j = \Delta_{ij} = x_i - x_j$$

for all solutions $x'_1, \ldots, x'_m$, $(y'_{ij})_{(i,j)\in\mathcal{A}}$ of (6). Hence $x'_j = x_j$. Since $j \in \bar{M}$, we can repeat the above argument with j in place of i and so on. This yields $x'_j = x_j$ for all $j \in \bar{M}$. Since the choice of the connected component was arbitrary, this shows that $x'_j = x_j$ for all $j \in M_{\mathcal{B}}$.

If $i \le m$ and $i \notin M_{\mathcal{B}}$, then $N_{\mathcal{B}}(i) = \emptyset$. This implies

$$\|x_i - x_j\|^2 < y_{ij} \quad \forall j \in N_{\mathcal{A}}(i).$$

Then, we can perturb x_i and obtain another solution $x'_1, \ldots, x'_m$, $(y_{ij})_{(i,j)\in\mathcal{A}}$ of (6) with $x'_i \neq x_i$. □

As a corollary of Proposition 5.1(c), we have that the solution of (6) is unique if and only if each connected component of the graph $(\{1, \ldots, n\}, \mathcal{B})$ contains an anchor index (i.e., $M_{\mathcal{B}} = \{1, \ldots, m\}$).

Proposition 5.1 shows that those points x_i with $i \in M_{\mathcal{B}}$ have the following three properties: (i) they satisfy (10), (ii) $\|x_i - x_j\|^2 = y_{ij}$ for all $j \in N_{\mathcal{B}}(i)$, and (iii) their positions are uniquely determined by the anchors $x_{m+1}, \ldots, x_n$ and $(y_{ij})_{(i,j)\in\mathcal{B}}$. Might the first two properties (i), (ii) imply property (iii)? This question is related to graph rigidity and uniqueness of graph realizability. However, the following example in Figure 5, suggested by Connelly [15], shows that this is not true. The outer three points are anchors, the edges of $\mathcal{B}$ are as shown, and the inner three points (sensors) form a triangle that can be twisted slightly clockwise or counterclockwise to be in two different positions, both of which have properties (i) and (ii).

FIG. 5. *An example in* $\Re^2$ *of nonunique sensor positions satisfying* (10) *and preserving distances.*

6. Analytic center solution of the SOCP relaxation. As mentioned in section 2, when we solve (7) using an interior-point method, the method will generally find not only an interior solution, but an analytic center of the solution set. We study this in more depth below. We first need the following lemma to relate the solutions of (6) and (7).

LEMMA 6.1. $(y_{ij})_{(i,j)\in\mathcal{A}}$ *is invariant over all solutions of* (6).

Proof. Let $\mathcal{B}$ be given by (9). Suppose we have two solutions of (6): $x_1, \ldots, x_m$, $(y_{ij})_{(i,j)\in\mathcal{A}}$ and $x'_1, \ldots, x'_m$, $(y'_{ij})_{(i,j)\in\mathcal{A}}$. Then, for each $(i,j) \in \mathcal{B}$, $y_{ij} = \|x_i - x_j\|^2$ and $y'_{ij} = \|x'_i - x'_j\|^2$ (with $x'_i = x_i$ for $i > m$), so that

$$\frac{y_{ij} + y'_{ij}}{2} = \left\| \frac{x_i - x_j}{2} + \frac{x'_i - x'_j}{2} \right\|^2 + \left\| \frac{x_i - x_j}{2} - \frac{x'_i - x'_j}{2} \right\|^2.$$

Since the solution set is convex so that $\frac{1}{2}(x_1 + x'_1), \ldots, \frac{1}{2}(x_m + x'_m)$, $(\frac{1}{2}(y_{ij} + y'_{ij}))_{(i,j)\in\mathcal{A}}$ also forms a solution of (6), the rightmost term must be zero, i.e., $x_i - x_j = x'_i - x'_j$. Thus $y_{ij} = y'_{ij}$.

For each $(i,j) \in \mathcal{A}\backslash\mathcal{B}$, we have $y_{ij} > \|x_i - x_j\|^2$ for all interior solutions $x_1, \ldots, x_m$, $(y_{ij})_{(i,j)\in\mathcal{A}}$ of (6), implying $y_{ij} = d_{ij}^2$. (If $y_{ij} \neq d_{ij}^2$, then y_{ij} can be perturbed to decrease $|y_{ij} - d_{ij}^2|$ and hence decrease the objective function value.) Taking closure yields that $y_{ij} = d_{ij}^2$ for all solutions $x_1, \ldots, x_m$, $(y_{ij})_{(i,j)\in\mathcal{A}}$ of (6) so that y_{ij} is unique. ∎

By using Lemma 6.1, we see that $(u_{ij})_{(i,j)\in\mathcal{A}}$ is invariant over all solutions of (7). Then, under Assumption 1, the limiting point of the central path for (7) would be an interior solution of (7) that maximizes (see [25, p. 223], [4, p. 384])

$$\sum_{(i,j)\in\mathcal{A}\backslash\mathcal{B}} \log\left(\left(y_{ij} + \frac{1}{4} \right)^2 - \left\| \left(y_{ij} - \frac{1}{4}, w_{ij} \right) \right\|^2 \right) = \sum_{(i,j)\in\mathcal{A}\backslash\mathcal{B}} \log\left(y_{ij} - \|x_i - x_j\|^2 \right).$$

Accordingly, we define an *analytic center solution* of (6) to be an interior solution of (6) that maximizes

$$\sum_{(i,j)\in\mathcal{A}\backslash\mathcal{B}} \log\left(y_{ij} - \|x_i - x_j\|^2 \right) \tag{13}$$

over all interior solutions. Thus, an analytic center solution in some sense maximizes the slacks $y_{ij} - \|x_i - x_j\|^2$ for all inactive constraints $(i,j) \in \mathcal{A} \setminus \mathcal{B}$. Its existence is guaranteed by Assumption 1. It is unique because of Proposition 5.1(c) and that, by Lemma 6.1 and the strict concavity of $\log(y_{ij} - \|\cdot\|^2)$, $x_i - x_j$ is unique for all $(i,j) \in \mathcal{A} \setminus \mathcal{B}$. This is the interior solution that a log-barrier interior-point method will likely find. If a barrier method based on a different barrier function is used to solve (7), then the interior solution found need not be the analytic center.

The next proposition verifies one of our observations from the example in section 2. This is further illustrated in Figure 7.

PROPOSITION 6.2. *If $x_1, \ldots, x_m$, $(y_{ij})_{(i,j)\in\mathcal{A}}$ is the analytic center solution of (6), then*

$$x_i \in \operatorname{conv}\{x_j\}_{j\in N_{\mathcal{A}}(i)}, \quad i = 1, \ldots, m.$$

Proof. We argue by contradiction. Suppose there exists $i \in \{1, \ldots, m\}$ such that

$$x_i \notin \operatorname{conv}\{x_j\}_{j\in N_{\mathcal{A}}(i)}.$$

Let p_i denote the nearest-point projection of x_i onto this convex hull. Then, as in the proof of Proposition 5.1(a), we have

$$\|p_i - x_j\|^2 < \|x_i - x_j\|^2 \quad \forall j \in N_{\mathcal{A}}(i).$$

Thus, replacing x_i by p_i would yield another interior solution of (6). Moreover, if $(i,j) \in \mathcal{A} \setminus \mathcal{B}$, then $y_{ij} - \|p_i - x_j\|^2 > y_{ij} - \|x_i - x_j\|^2 > 0$, so that

$$\log\left(y_{ij} - \|p_i - x_j\|^2\right) \; > \; \log\left(y_{ij} - \|x_i - x_j\|^2\right).$$

Summing these inequalities yields

$$\sum_{j \in N_{\mathcal{A}\setminus\mathcal{B}}(i)} \log\left(y_{ij} - \|p_i - x_j\|^2\right) > \sum_{j \in N_{\mathcal{A}\setminus\mathcal{B}}(i)} \log\left(y_{ij} - \|x_i - x_j\|^2\right).$$

This contradicts our assumption that $x_1, \ldots, x_m$, $(y_{ij})_{(i,j)\in\mathcal{A}}$ is the analytic center solution of (6). □

It is an open question whether a result analogous to Lemma 6.1 holds for the SDP relaxation (3); namely, is $\left(\langle b_{ij} b_{ij}^T, Z\rangle_F\right)_{(i,j)\in\mathcal{A}}$ invariant over all solutions of (3)? There does not appear to be a result analogous to Proposition 6.2 for SDP relaxation. In particular, (3) need not have any solution satisfying the convex hull condition of Proposition 6.2; see an example in Figure 1(a) of [31].

7. Error analysis for the SOCP relaxation. In practice, the distance d_{ij} has measurement error, i.e.,

$$d_{ij}^2 = \bar{y}_{ij} + \delta_{ij} \quad \forall (i,j) \in \mathcal{A},$$

where $\delta_{ij} \in \Re$ and $\bar{y}_{ij} = \|x_i^{\text{true}} - x_j^{\text{true}}\|^2$ for some $x_1^{\text{true}}, \ldots, x_m^{\text{true}}$ representing the true positions of the sensors, and with $x_i^{\text{true}} = x_i$ for $i > m$. What is the corresponding error in the solution of (6)? We study this question in this section.

In what follows, we denote for simplicity $x = (x_1, \ldots, x_m) \in \Re^d \times \cdots \times \Re^d$ and

$$q_{ij}(x) \stackrel{\text{def}}{=} \|x_i - x_j\|^2 - \bar{y}_{ij} \quad \forall (i,j) \in \mathcal{A}. \tag{14}$$

Also,

$$\Xi \stackrel{\text{def}}{=} \{x : q_{ij}(x) \le 0 \;\; \forall \; (i,j) \in \mathcal{A}\}. \tag{15}$$

Then Ξ contains the true solution $x^{\text{true}} = (x_1^{\text{true}}, \ldots, x_m^{\text{true}})$. By the convexity of q_{ij}, Ξ is a convex set and there exists $\bar{\mathcal{B}} \subseteq \mathcal{A}$ such that

$$q_{ij}(x) = 0 \; \forall x \in \Xi \quad \Longleftrightarrow \quad (i,j) \in \bar{\mathcal{B}}.$$

Since x^{true}, $(\bar{y}_{ij})_{(i,j)\in\mathcal{A}}$ is feasible for (6), any solution $x = (x_1, \ldots, x_m)$, $(y_{ij})_{(i,j)\in\mathcal{A}}$ of (6) satisfies

$$\sum_{(i,j)\in\mathcal{A}} |y_{ij} - d_{ij}^2| \le \sum_{(i,j)\in\mathcal{A}} |\bar{y}_{ij} - d_{ij}^2| = \sum_{(i,j)\in\mathcal{A}} |\delta_{ij}|. \tag{16}$$

Since $\|x_i - x_j\|^2 \le y_{ij}$ so that $q_{ij}(x) \le y_{ij} - \bar{y}_{ij}$, this yields

$$\begin{aligned}
\sum_{(i,j)\in\mathcal{A}} q_{ij}(x)_+ &\le \sum_{(i,j)\in\mathcal{A}} (y_{ij} - \bar{y}_{ij})_+ \\
&\le \sum_{(i,j)\in\mathcal{A}} |y_{ij} - \bar{y}_{ij}| \\
&\le \sum_{(i,j)\in\mathcal{A}} (|y_{ij} - d_{ij}^2| + |d_{ij}^2 - \bar{y}_{ij}|) \\
&\le 2 \sum_{(i,j)\in\mathcal{A}} |\delta_{ij}|, \qquad (17)
\end{aligned}$$

where $\alpha_+ \stackrel{\text{def}}{=} \max\{0, \alpha\}$.

Using Proposition 5.1, we show below that if the distance error is small so the right-hand side of (17) is small, then $(x_i)_{i\in M_{\mathcal{B}}}$ in a solution of (6) has small error (in fact, proportional to the square root of distance error), where $\mathcal{B}$ is given by (9); see Propositions 7.1 and 7.2. Moreover, we can find $\mathcal{B}$ from an interior solution of (6); see also section 9. Although there exist sensitivity analysis results for convex quadratic inequalities of the form (15), the results either make the restrictive assumption that Ξ has nonempty interior [23] or prove a much weaker result that the solution error is proportional to the $2^{|\mathcal{A}|+1}$th root of the distance error [33]; see discussions following Corollary 7.3. Existing sensitivity analysis results for general nonlinear programs make technical assumptions that either do not hold or are difficult to verify for (15); see, e.g., [12, sections 5.2, 5.3].

For any $\mathcal{B} \subseteq \mathcal{A}$,

$$\Xi_{\mathcal{B}} \stackrel{\text{def}}{=} \{x \in \Xi \ : \ q_{ij}(x) = 0 \ \forall (i,j) \in \mathcal{B}\}.$$

For any nonempty closed subset Ξ' of Ξ, let

$$\operatorname{dist}((x_1 \ldots, x_m), \Xi') \stackrel{\text{def}}{=} \min_{(\bar{x}_1,\ldots,\bar{x}_m)\in\Xi'} \max_{i=1,\ldots,m} \|x_i - \bar{x}_i\|.$$

PROPOSITION 7.1. (a) *For each $\epsilon > 0$, there exists a scalar $\delta > 0$ such that*

$$\Xi_{\mathcal{B}} \neq \emptyset \qquad \textit{and} \qquad \operatorname{dist}(x, \Xi_{\mathcal{B}}) \leq \epsilon$$

whenever $\mathcal{B}$ satisfies (9), *$x = (x_1, \ldots, x_m)$, $(y_{ij})_{(i,j)\in\mathcal{A}}$ is a solution of* (6), *and $\sum_{(i,j)\in\mathcal{A}} |\delta_{ij}| \leq \delta$.*

(b) *There exists an $\bar{\epsilon} > 0$ such that, for each $0 < \epsilon < \bar{\epsilon}$, there exists a scalar $\delta > 0$ such that*

$$\mathcal{B} \subseteq \bar{\mathcal{B}} \qquad \textit{and} \qquad \|x_i - x_i^{\text{true}}\| \leq \epsilon \quad \forall i \in M_{\mathcal{B}},$$

whenever $\mathcal{B}$ satisfies (9), *$x_1, \ldots, x_m$, $(y_{ij})_{(i,j)\in\mathcal{A}}$ is a solution of* (6), *and $\sum_{(i,j)\in\mathcal{A}} |\delta_{ij}| \leq \delta$.*

Proof. (a) Fix any $\epsilon > 0$. If the desired δ does not exist, there would exist $(\delta_{ij}^t)_{(i,j)\in\mathcal{A}}$, $t = 1, 2, \ldots$, with $\sum_{(i,j)\in\mathcal{A}} |\delta_{ij}^t| \to 0$, and a $\mathcal{B} \subseteq \mathcal{A}$ satisfying (9) with $d_{ij}^2 = \bar{y}_{ij} + \delta_{ij}^t$ for $(i,j) \in \mathcal{A}$ in (6), $t = 1, 2, \ldots$. In addition, for each $t = 1, 2, \ldots$, there would exist a solution $x^t = (x_1^t, \ldots, x_m^t)$, $y^t = (y_{ij}^t)_{(i,j)\in\mathcal{A}}$, of (6) with $d_{ij}^2 = \bar{y}_{ij} + \delta_{ij}^t$, and yet either the set $\Xi_{\mathcal{B}}$ is empty or $\operatorname{dist}(x^t, \Xi_{\mathcal{B}}) > \epsilon$ for all t.

We see from (16) that $\{y_{ij}^t\} \to \bar{y}_{ij}$ for all $(i,j) \in \mathcal{A}$. Also, we can assume without loss of generality that $\{x^t\}$ is bounded.[5] By passing to a subsequence if necessary, we assume that $\{x^t\}$ converges to some $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_m)$. By (17),

$$\sum_{(i,j)\in\mathcal{A}} q_{ij}(x^t)_+ \leq 2 \sum_{(i,j)\in\mathcal{A}} |\delta_{ij}^t|, \quad t = 1, 2, \ldots.$$

[5] Consider any connected component $\mathcal{C}$ of the graph $\mathcal{G} = (\{1, \ldots, n\}, \mathcal{A})$. If $\mathcal{C}$ contains an anchor index, then $\{x_i^t\}$ is bounded for all $i \leq m$ in $\mathcal{C}$. If $\mathcal{C}$ does not contain an anchor index, then it can be seen that $\{\|x_i^t - x_j^t\|\}$ is bounded for all i and j in $\mathcal{C}$, so we can translate x_i^t for all i in $\mathcal{C}$ by the same displacement (thus preserving the distances between them) so that one of them is at the origin.

This yields in the limit $\sum_{(i,j)\in\mathcal{A}} q_{ij}(\bar{x})_+ \le 0$, implying $\bar{x} \in \Xi$. Since $\mathcal{B}$ satisfies (9) with $d_{ij}^2 = \bar{y}_{ij} + \delta_{ij}^t$ for $(i,j) \in \mathcal{A}$ in (6), we also have

$$\|x_i^t - x_j^t\|^2 = y_{ij}^t \quad \forall (i,j) \in \mathcal{B},\ t = 1,2,\ldots \quad (\text{with } x_i^t = x_i\ \forall i > m).$$

This yields in the limit

$$\|\bar{x}_i - \bar{x}_j\|^2 = \bar{y}_{ij} \quad \forall (i,j) \in \mathcal{B} \quad (\text{with } \bar{x}_i = x_i\ \forall i > m),$$

implying $\bar{x} \in \Xi_{\mathcal{B}}$. Moreover $\max_{i=1,\ldots,m} \|x_i^t - \bar{x}_i\| \to 0$. This contradicts $\Xi_{\mathcal{B}} = \emptyset$ or $\operatorname{dist}(x^t, \Xi_{\mathcal{B}}) > \epsilon$ for all t.

(b) Since each q_{ij} is convex, Ξ has an interior solution, i.e., $x' = (x_1', \ldots, x_m') \in \Xi$ satisfying

$$\|x_i' - x_j'\|^2 < \bar{y}_{ij} \quad \forall (i,j) \in \mathcal{A} \setminus \bar{\mathcal{B}}, \tag{18}$$

where we let $x_i' = x_i$ for $i > m$.

By (a), for any $\epsilon > 0$, there exists $\delta > 0$ such that, for any interior solution $x_1, \ldots, x_m$, $(y_{ij})_{(i,j)\in\mathcal{A}}$ of (6) with $\sum_{(i,j)\in\mathcal{A}} |\delta_{ij}| \le \delta$, there exists $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_m) \in \Xi_{\mathcal{B}}$ with $\max_{i=1,\ldots,m} \|x_i - \bar{x}_i\| \le \epsilon$, where $\mathcal{B}$ satisfies (9). Let

$$\begin{aligned} d_i &= \bar{x}_i - x_i, \quad i = 1, \ldots, n, \\ d_i' &= x_i' - \bar{x}_i, \quad i = 1, \ldots, n, \end{aligned}$$

with $\bar{x}_i = x_i$ for $i > m$. Since $\bar{x}$ and x' are both in Ξ, we have $\|\bar{x}_i - \bar{x}_j\|^2 = \|x_i' - x_j'\|^2$ for all $(i,j) \in \bar{\mathcal{B}}$, which yields (also see the proof of Lemma 6.1)

$$d_i' - d_j' = 0 \quad \forall (i,j) \in \bar{\mathcal{B}}. \tag{19}$$

For each $(i,j) \in \mathcal{B}$, we have

$$\|x_i' - x_j'\|^2 = \bar{y}_{ij} - s_{ij}$$

for some $s_{ij} \ge 0$, or, equivalently,

$$\|d_i' - d_j' + \bar{x}_i - \bar{x}_j\|^2 = \|\bar{x}_i - \bar{x}_j\|^2 - s_{ij}.$$

Expanding the quadratics yields

$$2(\bar{x}_i - \bar{x}_j)^T (d_i' - d_j') = -\|d_i' - d_j'\|^2 - s_{ij}$$

or, equivalently,

$$\begin{aligned} 2(x_i - x_j)^T (d_i' - d_j') &= -\|d_i' - d_j'\|^2 - s_{ij} - 2(d_i - d_j)^T (d_i' - d_j') \\ &\le -\|d_i' - d_j'\|^2 - s_{ij} + 2\|d_i - d_j\|\|d_i' - d_j'\|. \end{aligned}$$

Since $s_{ij} \ge 0$, this implies that

$$(x_i - x_j)^T (d_i' - d_j') < 0 \quad \text{whenever} \quad \|d_i - d_j\| < \|d_i' - d_j'\|/2. \tag{20}$$

Let us choose

$$\epsilon < \frac{1}{4} \min_{(i,j)\in\mathcal{A}\setminus\bar{\mathcal{B}}} \{\sqrt{\bar{y}_{ij}} - \|x_i' - x_j'\|\}, \tag{21}$$

where the right-hand side is positive by (18). For each $(i,j) \in \mathcal{B}$ with $d'_i - d'_j \neq 0$, we have from (19) that $(i,j) \notin \bar{\mathcal{B}}$. Then, by using $\|d_i\| \leq \epsilon$ and (21), we have

$$\|d_i - d_j\| \leq 2\epsilon < \frac{\sqrt{\bar{y}_{ij}} - \|x'_i - x'_j\|}{2} \leq \frac{\|d'_i - d'_j\|}{2},$$

where the last inequality follows from

$$\|d'_i - d'_j\| = \|x'_i - x'_j - (\bar{x}_i - \bar{x}_j)\| \geq \|\bar{x}_i - \bar{x}_j\| - \|x'_i - x'_j\| = \sqrt{\bar{y}_{ij}} - \|x'_i - x'_j\|.$$

Then, by (20), $(x_i - x_j)^T(d'_i - d'_j) < 0$. Thus, for each $(i,j) \in \mathcal{B}$ and for all $\alpha > 0$ sufficiently small, we have

$$\|(x_i + \alpha d'_i) - (x_j + \alpha d'_j)\|^2 \begin{cases} < \|x_i - x_j\|^2 = y_{ij} & \text{if } d'_i - d'_j \neq 0; \\ = \|x_i - x_j\|^2 = y_{ij} & \text{if } d'_i - d'_j = 0. \end{cases}$$

Also, for each $(i,j) \in \mathcal{A} \setminus \mathcal{B}$, since $\|x_i - x_j\|^2 < y_{ij}$, we have

$$\|(x_i + \alpha d'_i) - (x_j + \alpha d'_j)\|^2 < y_{ij}$$

for all $\alpha > 0$ sufficiently small. Thus, if $d'_i - d'_j \neq 0$ for some $(i,j) \in \mathcal{B}$, this would contradict the definition of $\mathcal{B}$. Hence $d'_i - d'_j = 0$ for all $(i,j) \in \mathcal{B}$. This in turn implies

$$\|x'_i - x'_j\|^2 = \|\bar{x}_i - \bar{x}_j\|^2 = \bar{y}_{ij} \quad \forall (i,j) \in \mathcal{B}.$$

Hence (18) yields that $\mathcal{B} \subseteq \bar{\mathcal{B}}$.

Since $\bar{x} \in \Xi$, by applying Proposition 5.1(c) we have

$$\bar{x}_i = x_i^{\text{true}} \quad \forall i \in M_{\bar{\mathcal{B}}}.$$

Since $\mathcal{B} \subseteq \bar{\mathcal{B}}$, $M_{\mathcal{B}} \subseteq M_{\bar{\mathcal{B}}}$. Thus

$$\|x_i - x_i^{\text{true}}\| = \|x_i - \bar{x}_i\| \leq \epsilon \quad \forall i \in M_{\mathcal{B}}. \qquad \square$$

From the proof of Proposition 7.1(b) we see that we can take $\bar{\epsilon}$ to be the right-hand side of (21), maximized over all $(x'_1, \ldots, x'_m) \in \Xi$. Proposition 7.1(b) says that if the distance error is not too large, then the error in the position of those sensors indexed by $M_{\mathcal{B}}$ is also not too large. However, it does not say how fast the position error grows with the distance error. We show below that the position error grows at most like the square root of the distance error.

We say that $\mathcal{B} \subseteq \bar{\mathcal{B}}$ is *active with respect to* $\mathcal{M} \subseteq \{1, \ldots, m\}$ if

$$q_{ij}(x) \leq 0 \ \ \forall (i,j) \in \mathcal{B}, \ \ x_i = x_i^{\text{true}} \ \ \forall i \notin \mathcal{M} \quad \Longrightarrow \quad q_{ij}(x) = 0 \ \ \forall (i,j) \in \mathcal{B}.$$

We say that $\mathcal{B}$ is *minimally active with respect to* $\mathcal{M}$ if there is no proper subset of $\mathcal{B}$ that is active with respect to $\mathcal{M}$.

PROPOSITION 7.2. *There exists a constant $K > 0$ such that*

$$\max_{i \in M_{\bar{\mathcal{B}}}} \|x_i - x_i^{\text{true}}\| \leq K \max_{(i,j) \in \bar{\mathcal{B}}} q_{ij}(x)_+^{1/2} \quad \forall x = (x_1, \ldots, x_m).$$

Proof. If $\bar{\mathcal{B}} = \emptyset$, then our proof is complete. Otherwise, by its definition, $\bar{\mathcal{B}}$ is active with respect to $\{1, \ldots, m\}$. Then, there exists nonempty $\mathcal{B}_1 \subseteq \bar{\mathcal{B}}$ that is

minimally active with respect to $\{1,\ldots,m\}$. By using Gordan's theorem as in the proof of [33, Theorem 3.1], there exist $\lambda_{ij} > 0$, $(i,j) \in \mathcal{B}_1$, satisfying

$$\sum_{(i,j)\in\mathcal{B}_1} \nabla q_{ij}(x^{\text{true}})\lambda_{ij} = 0.^6 \tag{22}$$

Fix any $x = (x_1,\ldots,x_m)$. For each $(i,j) \in \mathcal{B}_1$, we have from $q_{ij}(x^{\text{true}}) = 0$ that

$$q_{ij}(x) = \nabla q_{ij}(x^{\text{true}})^T(x - x^{\text{true}}) + \|x_i - x_j - (x_i^{\text{true}} - x_j^{\text{true}})\|^2.$$

Multiplying both sides by λ_{ij} and summing over all $(i,j) \in \mathcal{B}_1$ and using (22) yield

$$\sum_{(i,j)\in\mathcal{B}_1} q_{ij}(x)\lambda_{ij} = \sum_{(i,j)\in\mathcal{B}_1} \|x_i - x_j - (x_i^{\text{true}} - x_j^{\text{true}})\|^2\lambda_{ij}.$$

Thus

$$\|x_i - x_j - (x_i^{\text{true}} - x_j^{\text{true}})\|^2\lambda_{ij} \le \sum_{(i,j)\in\mathcal{B}_1} \lambda_{ij} \cdot \max_{(i,j)\in\mathcal{B}_1} q_{ij}(x)_+ \quad \forall (i,j) \in \mathcal{B}_1.$$

This in turn implies

$$\|x_i - x_j - (x_i^{\text{true}} - x_j^{\text{true}})\| \le C_1 \max_{(i,j)\in\mathcal{B}_1} q_{ij}(x)_+^{1/2} \quad \forall (i,j) \in \mathcal{B}_1, \tag{23}$$

where

$$C_1 \stackrel{\text{def}}{=} \left(\frac{\sum_{(i,j)\in\mathcal{B}_1}\lambda_{ij}}{\min_{(i,j)\in\mathcal{B}_1}\lambda_{ij}}\right)^{1/2}.$$

We can then apply Proposition 5.1(b) with d_{ij}, $\mathcal{A}$, $\mathcal{B}$, $\{1,\ldots,m\}$ replaced by, respectively, $\sqrt{\bar{y}_{ij}}$, $\mathcal{B}_1$, $\mathcal{B}_1$, $\mathcal{M}_1 \stackrel{\text{def}}{=} \{i \in \{1,\ldots,m\} : N_{\mathcal{B}_1}(i) \neq \emptyset\} = M_{\mathcal{B}_1}$. This yields that each connected component of the graph $\mathcal{G}_1 \stackrel{\text{def}}{=} (\mathcal{M}_1 \cup \{m+1,\ldots,n\}, \mathcal{B}_1)$ contains an anchor index $j \in \{m+1,\ldots,n\}$. (In fact, this graph is connected since $\mathcal{B}_1$ is minimally active with respect to $\mathcal{M}_1$.) Then, for each $i \in N_{\mathcal{B}_1}(j)$, we have from (23) and $x_j = x_j^{\text{true}}$ that

$$\|x_i - x_i^{\text{true}}\| \le C_1 \max_{(i,j)\in\mathcal{B}_1} q_{ij}(x)_+^{1/2}.$$

Continuing this argument with each neighbor of i in $\mathcal{G}_1$, and so on, we obtain that

$$\|x_i - x_i^{\text{true}}\| \le C_1 D_1 \max_{(i,j)\in\mathcal{B}_1} q_{ij}(x)_+^{1/2} \quad \forall i \in \mathcal{M}_1, \tag{24}$$

[6] Why? Since $\mathcal{B}_1$ is active with respect to $\{1,\ldots,m\}$ and $q_{ij}(x^{\text{true}}) = 0$ for all $(i,j) \in \mathcal{B}_1$, the linear system $\nabla q_{ij}(x^{\text{true}})^T d < 0$, $(i,j) \in \mathcal{B}_1$, is infeasible. By Gordan's theorem [13, p. 23], there exist $\lambda_{ij} \ge 0$ for $(i,j) \in \mathcal{B}_1$, not all zero, satisfying (22). Let $\hat{\mathcal{B}}_1 \stackrel{\text{def}}{=} \{(i,j) \in \mathcal{B}_1 : \lambda_{ij} > 0\}$. By Gordan's theorem again, the linear system $\nabla q_{ij}(x^{\text{true}})^T d < 0$, $(i,j) \in \hat{\mathcal{B}}_1$, is infeasible. If $\hat{\mathcal{B}}_1 \neq \mathcal{B}_1$, then the quadratic system $q_{ij}(x) < 0$, $(i,j) \in \hat{\mathcal{B}}_1$, would be feasible. (Otherwise there would exist a nonempty $\tilde{\mathcal{B}}_1 \subseteq \hat{\mathcal{B}}_1$ such that $q_{ij}(x) = 0$, $(i,j) \in \tilde{\mathcal{B}}_1$, whenever $q_{ij}(x) \le 0$, $(i,j) \in \hat{\mathcal{B}}_1$. Choose $\tilde{\mathcal{B}}_1$ to be maximal. Then $\tilde{\mathcal{B}}_1$ would be active with respect to $\{1,\ldots,m\}$, contradicting $\mathcal{B}_1$ being minimally active with respect to $\{1,\ldots,m\}$.) Then the linear system $\nabla q_{ij}(x^{\text{true}})^T d < 0$, $(i,j) \in \hat{\mathcal{B}}_1$, would be feasible, which is a contradiction. Thus $\hat{\mathcal{B}}_1 = \mathcal{B}_1$.

where $D_1 \stackrel{\text{def}}{=} \max_{i\in\mathcal{M}_1} \min_{j\notin\mathcal{M}_1}$(minimum number of edges in a path between i and j in $\mathcal{G}_1$).

If $\bar{\mathcal{B}} = \mathcal{B}_1$, then our proof is complete. Otherwise, $\bar{\mathcal{B}} \setminus \mathcal{B}_1$ is active with respect to $\{1,\dots,m\} \setminus \mathcal{M}_1$. Then, there exists nonempty $\mathcal{B}_2 \subseteq \bar{\mathcal{B}} \setminus \mathcal{B}_1$ that is minimally active with respect to $\{1,\dots,m\} \setminus \mathcal{M}_1$. Repeating the above argument, we obtain that

$$\|x_i - x_j - (x_i^{\text{true}} - x_j^{\text{true}})\| \le C_2 \max_{(i,j)\in\mathcal{B}_2} q_{ij}(x)_+^{1/2} \quad \forall (i,j) \in \mathcal{B}_2, \tag{25}$$

with C_2 defined analogously as C_1 and with $x_i = x_i^{\text{true}}$ for $i \in \mathcal{M}_1$. We can then apply Proposition 5.1(b) with d_{ij}, $\mathcal{A}$, $\mathcal{B}$, $\{1,\dots,m\}$ replaced by, respectively, $\sqrt{\bar{y}_{ij}}$, $\mathcal{B}_2$, $\mathcal{B}_2$, $\mathcal{M}_2 \stackrel{\text{def}}{=} \{i \in \{1,\dots,m\} \setminus \mathcal{M}_1 : N_{\mathcal{B}_2}(i) \neq \emptyset\}$. This yields that each connected component of the graph $\mathcal{G}_2 \stackrel{\text{def}}{=} (\mathcal{M}_1 \cup \mathcal{M}_2 \cup \{m+1,\dots,n\}, \mathcal{B}_2)$ contains a node $j \in \mathcal{M}_1 \cup \{m+1,\dots,n\}$. Then, for each $i \in N_{\mathcal{B}_2}(j)$, we have from (25) and (24) that

$$\|x_i - x_i^{\text{true}}\| \le C_1 D_1 \max_{(i,j)\in\mathcal{B}_1} q_{ij}(x)_+^{1/2} + C_2 \max_{(i,j)\in\mathcal{B}_2} q_{ij}(x)_+^{1/2}.$$

Continuing this argument with each neighbor of i in $\mathcal{G}_2$, and so on, we obtain that

$$\|x_i - x_i^{\text{true}}\| \le C_1 D_1 \max_{(i,j)\in\mathcal{B}_1} q_{ij}(x)_+^{1/2} + C_2 D_2 \max_{(i,j)\in\mathcal{B}_2} q_{ij}(x)_+^{1/2} \quad \forall i \in \mathcal{M}_2,$$

with D_2 defined analogously as D_1.

Continuing the above argument inductively completes the proof. □

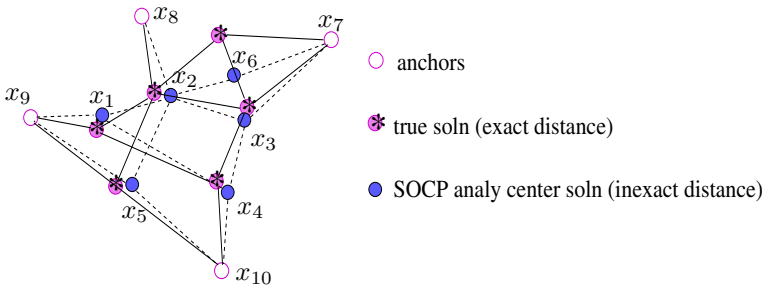


FIG. 6. *In this example,* $\mathcal{M}_1 = \{1,2,3,4\}$, $\mathcal{B}_1 = \{(1,2),(1,4),(1,9),(2,3),(2,8),(3,7),(3,4),(4,10)\}$, $\mathcal{M}_2 = \{5\}$, $\mathcal{B}_2 = \{(5,2),(5,9),(5,10)\}$, *and* $\bar{\mathcal{B}} = \mathcal{B}_1 \cup \mathcal{B}_2$. *Removing a point indexed by* $\mathcal{M}_1$ *affects points indexed by* $\mathcal{M}_2$ *but not conversely.*

The proof of Proposition 7.2 shows that the points indexed by $\mathcal{M}_1$, which are the sensors "nearest" to the anchors, are the least sensitive to distance measurement errors. An important (and intuitively reasonable) result shown by Proposition 7.2 is that the errors affect the sensor positions additively as they percolate to $\mathcal{M}_2$, and so on; see Figure 6 for an illustrative example.

COROLLARY 7.3. *There exists a constant* $L > 0$ *such that*

$$\operatorname{dist}(x,\Xi) \le L \max_{(i,j)\in\mathcal{A}\setminus\bar{\mathcal{B}}} q_{ij}(x)_+ + LK \max_{(i,j)\in\bar{\mathcal{B}}} q_{ij}(x)_+^{1/2} \quad \forall x = (x_1,\dots,x_m),$$

where K *is defined as in Proposition* 7.2.

Proof. Consider the system of convex quadratic inequalities and linear equations in $x = (x_1,\dots,x_m)$:

$$q_{ij}(x) \le 0 \ \ \forall (i,j) \in \mathcal{A} \setminus \bar{\mathcal{B}}, \qquad x_i = x_i^{\text{true}} \ \ \forall i \in M_{\bar{\mathcal{B}}}.$$

By applying Proposition 5.1(c) with d_{ij} replaced by $\sqrt{\bar{y}_{ij}}$, we see that Ξ equals the solution set of this system. Moreover, each interior solution of Ξ satisfies the quadratic inequalities strictly. Thus, applying a result of Luo and Luo [23], there exists $L > 0$ such that

$$\mathrm{dist}(x, \Xi) \le L \max_{(i,j)\in\mathcal{A}\setminus\bar{\mathcal{B}}} q_{ij}(x)_+ + L \max_{i\in M_{\bar{\mathcal{B}}}} \|x_i - x_i^{\text{true}}\| \quad \forall x = (x_1, \ldots, x_m).$$

Using Proposition 7.2 to bound the second term on the right-hand side completes the proof. □

The error bound in Corollary 7.3 sharpens the Hölderian error bound of Wang and Pang [33] for general convex quadratic inequalities. In particular, a direct application of the result in [33] yields the existence of $\tau > 0$ and integer $\ell \le |\mathcal{A}| + 1$ such that

$$\mathrm{dist}(x, \Xi) \le \kappa \max_{(i,j)\in\mathcal{A}} \left(q_{ij}(x)_+ + q_{ij}(x)_+^{1/2^\ell} \right) \quad \forall x = (x_1, \ldots, x_m).$$

An example in [33] shows that, for general convex quadratic functions q_{ij}, $\ell = |\mathcal{A}|$ is possible. Corollary 7.3 in effect shows that we can take $\ell = 1$ in the special case where each q_{ij} has the form (14). It is an open question whether the active set index $\bar{\mathcal{B}}$ can be identified using the Lipschitzian error bound. The difficulty lies in that $\bar{y}_{ij}$ is unknown, so that $q_{ij}(x)$ cannot be directly evaluated.

Last, we show that the x-component of the analytic center solution of (6) converges to the analytic center solution of Ξ as the distance error goes to zero.

PROPOSITION 7.4. *Under Assumption* 1, *let* $x^c = (x_1^c, \ldots, x_m^c)$, $(y_{ij}^c)_{(i,j)\in\mathcal{A}}$ *be the analytic center solution of* (6). *As* $\sum_{(i,j)\in\mathcal{A}} |\delta_{ij}| \to 0$, x^c *converges to the analytic center* $\bar{x}^c = (\bar{x}_1^c, \ldots, \bar{x}_m^c)$ *of* Ξ.

Proof. For $i = 1, \ldots, m$, let

$$\tilde{x}_i \stackrel{\text{def}}{=} \begin{cases} x_i^c & \text{if } i \in M_{\bar{\mathcal{B}}}; \\ \bar{x}_i^c & \text{if } i \notin M_{\bar{\mathcal{B}}}. \end{cases}$$

By $\bar{x}^c \in \Xi$ and Proposition 7.2, we have $\bar{x}_i^c = x_i^{\text{true}}$ for all $i \in M_{\bar{\mathcal{B}}}$. Let

$$\rho \stackrel{\text{def}}{=} - \max_{(i,j)\in\mathcal{A}\setminus\bar{\mathcal{B}}} q_{ij}(\bar{x}^c) > 0. \tag{26}$$

Suppose $\sum_{(i,j)\in\mathcal{A}} |\delta_{ij}| \le \delta$ for some $\delta > 0$. By (16), (17), and Proposition 7.2, we have

$$\max_{(i,j)\in\mathcal{A}} |y_{ij}^c - \bar{y}_{ij}| \le \delta, \qquad \max_{i\in M_{\bar{\mathcal{B}}}} \|x_i^c - \bar{x}_i^c\| \le K\sqrt{2\delta}. \tag{27}$$

For each $(i, j) \in \mathcal{A}$, consider the following three cases: (i) If $i \in M_{\bar{\mathcal{B}}}$ and $j \in M_{\bar{\mathcal{B}}}$, then

$$\|\tilde{x}_i - \tilde{x}_j\|^2 = \|x_i^c - x_j^c\|^2 \le y_{ij}^c.$$

(ii) If $i \notin M_{\bar{\mathcal{B}}}$ and $j \notin M_{\bar{\mathcal{B}}}$, then $(i, j) \notin \bar{\mathcal{B}}$ and hence (26), (27) yield

$$\|\tilde{x}_i - \tilde{x}_j\|^2 = \|\bar{x}_i^c - \bar{x}_j^c\|^2 \le \bar{y}_{ij} - \rho \le y_{ij}^c + \delta - \rho.$$

(iii) If $i \in M_{\bar{\mathcal{B}}}$ and $j \notin M_{\bar{\mathcal{B}}}$, then $(i,j) \notin \bar{\mathcal{B}}$ and hence (26), (27) yield

$$\begin{aligned}\|\tilde{x}_i - \tilde{x}_j\|^2 &= \|x_i^c - \bar{x}_j^c\|^2 \\ &\le (\|x_i^c - \bar{x}_i^c\| + \|\bar{x}_i^c - \bar{x}_j^c\|)^2 \\ &\le (K\sqrt{2\delta} + \|\bar{x}_i^c - \bar{x}_j^c\|)^2 \\ &\le 2K^2\delta + 2K\sqrt{2\delta}\sqrt{\bar{y}_{ij}} + \bar{y}_{ij} - \rho \\ &\le 2K^2\delta + 2K\sqrt{2\delta}\sqrt{\bar{y}_{ij}} + y_{ij}^c + \delta - \rho.\end{aligned}$$

Notice that $(i,j) = (j,i)$, so the case of $i \notin M_{\bar{\mathcal{B}}}$ and $j \in M_{\bar{\mathcal{B}}}$ is covered by case (iii). Since $x^c = (x_1^c, \ldots, x_m^c)$, $(y_{ij}^c)_{(i,j)\in\mathcal{A}}$ is a solution of (6) and $\rho > 0$, the above analysis shows that, for δ sufficiently small, $\tilde{x}_1, \ldots, \tilde{x}_m$, $(y_{ij}^c)_{(i,j)\in\mathcal{A}}$ is an interior solution of (6). Since $x^c = (x_1^c, \ldots, x_m^c)$, $(y_{ij}^c)_{(i,j)\in\mathcal{A}}$ is the analytic center solution of (6), this implies

$$\sum_{(i,j)\in\mathcal{A}\setminus\mathcal{B}} \log\left(y_{ij}^c - \|x_i^c - x_j^c\|^2\right) \ge \sum_{(i,j)\in\mathcal{A}\setminus\mathcal{B}} \log\left(y_{ij}^c - \|\tilde{x}_i - \tilde{x}_j\|^2\right),$$

where $\mathcal{B}$ satisfies (9) (with $x_i^c = \tilde{x}_i = x_i$ for $i > m$). By Proposition 7.1(b), we have $\mathcal{B} \subseteq \bar{\mathcal{B}}$ for δ sufficiently small. For $(i,j) \in \bar{\mathcal{B}}$, since $i \in M_{\bar{\mathcal{B}}}$ and $j \in M_{\bar{\mathcal{B}}}$, we have $\|\tilde{x}_i - \tilde{x}_j\|^2 = \|x_i^c - x_j^c\|^2$. Thus we further have

$$\sum_{(i,j)\in\mathcal{A}\setminus\bar{\mathcal{B}}} \log\left(y_{ij}^c - \|x_i^c - x_j^c\|^2\right) \ge \sum_{(i,j)\in\mathcal{A}\setminus\bar{\mathcal{B}}} \log\left(y_{ij}^c - \|\tilde{x}_i - \tilde{x}_j\|^2\right).$$

Taking $\delta \to 0$, we have from (27) that $y_{ij}^c \to \bar{y}_{ij}$ for all $(i,j) \in \mathcal{A}\setminus\bar{\mathcal{B}}$ and $\tilde{x}_i = x_i^c \to \bar{x}_i^c$ for all $i \in M_{\bar{\mathcal{B}}}$. Also, $\tilde{x}_i = \bar{x}_i^c$ for all $i \notin M_{\bar{\mathcal{B}}}$. Thus, we obtain in the limit that any cluster point $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_m)$ of x^c (which exists since x^c is uniformly bounded by Assumption 1) belongs to Ξ (using (17) and Corollary 7.3) and satisfies

$$\sum_{(i,j)\in\mathcal{A}\setminus\bar{\mathcal{B}}} \log\left(\bar{y}_{ij} - \|\bar{x}_i - \bar{x}_j\|^2\right) \ge \sum_{(i,j)\in\mathcal{A}\setminus\bar{\mathcal{B}}} \log\left(\bar{y}_{ij} - \|\bar{x}_i^c - \bar{x}_j^c\|^2\right)$$

(with $\bar{x}_i = \bar{x}_i^c = x_i$ for $i > m$). This shows that $\bar{x}$ is an analytic center of Ξ, so that in fact $\bar{x} = \bar{x}^c$. □

It is an open question whether the results of this section extend to the SDP relaxation (3).

8. Methods for solving the SOCP relaxation. We saw in previous sections that the SOCP relaxation (6), though weaker than the SDP relaxation (3), has the advantage of a smaller problem size and its interior solutions are useful for identifying sensors that are accurately positioned. What method would best solve (6) and, in particular, find an interior solution? A primal-dual interior-point method can find an analytic center solution of SOCP with good accuracy. However, as we will see in section 9, applying an interior-point method directly to (7) can be slow, due to the large size of the SOCP. We tried adapting the distributed SDP method of Biswas and Ye [11] to the SOCP relaxation. However, possibly due to the weaker SOCP relaxation, the resulting distributed SOCP method was not satisfactory. Further studies are needed. Below we describe a third method, based on smoothing and (block) coordinate gradient descent, which can find an interior solution faster, as we will see in section 9. This method has the nice feature that its computations easily distribute over many processors in parallel.

First, we observe that, for any $d \in \Re$,

$$\min_{y \geq z} |y - d^2| = [z - d^2]_+ \quad \forall z \in \Re,$$

where $[t]_+ = \max\{0, t\}$. Thus, we can rewrite the SOCP relaxation (6) as the unconstrained optimization problem

$$v_{\text{socp}} = \min_{x_1,\ldots,x_m} \sum_{(i,j)\in\mathcal{A}} [\|x_i - x_j\|^2 - d_{ij}^2]_+. \tag{28}$$

The objective function is convex, but nonsmooth due to the term $\max\{0, \cdot\}$. It is well known in the context of complementarity problems that a smoothing approach can be effective in handling this type of nonsmoothness; see [14, 18] and references therein. In particular, for any function $h : \Re \to \Re$ that is smooth and convex and satisfies $\lim_{t\to-\infty} h(t) = \lim_{t\to\infty} h(t) - t = 0$, we have that

$$\lim_{\mu\downarrow 0} \mu h(t/\mu) \;=\; [t]_+.$$

Thus, for $\mu > 0$ and small, we have $\mu h(t/\mu) \approx [t]_+$. In our numerical tests, we use a popular choice of h due to Chen, Harker, Kanzow, and Smale:

$$h(t) = ((t^2 + 4)^{1/2} + t)/2.$$

Thus, the nonsmooth problem (28) is approximated by the smooth problem, parameterized by $\mu > 0$:

$$\min_{x_1,\ldots,x_m} \sum_{(i,j)\in\mathcal{A}} \mu h\left(\frac{\|x_i - x_j\|^2 - d_{ij}^2}{\mu}\right). \tag{29}$$

For each $\mu > 0$, the objective function is smooth and convex and, as $\mu \to 0$, any cluster point of the solution of (29) is a solution of (28).

Since we wish to find an interior solution, following the interior-point approach, we add a log-barrier term and consider

$$\min_{y\geq z} |y - d^2| - \mu \log(y - z) = [z + \mu - d^2]_+ - \mu \log\left(\mu + [d^2 - z - \mu]_+\right) \quad \forall z \in \Re.$$

This is a convex function of z. Upon smoothing $[\cdot]_+$ by $\mu h(\cdot/\mu)$, we obtain the corresponding smooth barrier problem:

$$\min_{x=(x_1,\ldots,x_m)} f_\mu(x) \;\stackrel{\text{def}}{=}\; \sum_{(i,j)\in\mathcal{A}} \mu h\left(\frac{t_{ij}}{\mu}\right) - \mu \log\left(1 + h\left(\frac{-t_{ij}}{\mu}\right)\right)_{t_{ij}=\|x_i-x_j\|^2+\mu-d_{ij}^2}. \tag{30}$$

Here, for simplicity, we used the same parameter μ for the log-barrier and the smoothing function. Notice that the objective function f_μ is partially separable, being a sum of functions each of which depends only on the difference of neighboring points. This suggests that a block-coordinate descent approach may be efficient for solving (30), whereby at each iteration the objective function f_μ is minimized with respect to x_i, for some $i \in \{1, \ldots, m\}$, while the other points are held fixed at their current value. Since exact minimization is expensive, the minimization is done only inexactly. In

particular, we minimize a quadratic approximation of f_μ with respect to x_i to generate the descent direction d_i and then minimize f_μ inexactly along d_i using an Armijo stepsize rule [6]. We decrease μ whenever $\|\nabla f_\mu(x)\|$ is small relative to μ. The method, which we refer to as the smoothing coordinate gradient descent (SCGD) method, is described more precisely below.

0. Initialize $\mu > 0$ and $x = (x_1, \ldots, x_m)$. Choose $\mu^{\text{final}} > 0$ and a continuous function $\psi : (0, \infty) \to (0, \infty)$ satisfying $\lim_{\mu \downarrow 0} \psi(\mu) = 0$. Choose stepsize parameters $0 < \beta < 1$, $0 < \sigma < \frac{1}{2}$. Go to step 1.
1. If there exists an $i \in \{1, \ldots, m\}$ satisfying $\|\nabla_{x_i} f_\mu(x)\| > \psi(\mu)$, then set

$$d_i = -[H_i]^{-1} \nabla_{x_i} f_\mu(x),$$

update

$$x_i^{\text{new}} = x_i + \alpha d_i,$$

and repeat step 1, where $H_i \in \Re^{d \times d}$ is a user-chosen symmetric positive definite matrix, and α is the largest element of $\{1, \beta, (\beta)^2, \ldots\}$ satisfying

$$f_\mu(x_1, \ldots, x_i + \alpha d_i, \ldots, x_m) \leq f_\mu(x) - \alpha \sigma d_i^T \nabla_{x_i} f_\mu(x).$$

Otherwise, go to step 2.
2. If $\mu \leq \mu^{\text{final}}$, then stop. Otherwise decrease μ, and return to step 1.

The SCGD method is highly parallelizable since updating x_i requires knowledge of only neighboring points $\{x_j\}_{j \in N_{\mathcal{A}}(i)}$, so nonneighbors can update their positions simultaneously. Thus the computation can be distributed over the sensors, with each sensor communicating with its neighbors only.

In our current implementation of the SCGD method, we choose

$$H_i = \nabla^2_{x_i x_i} f_\mu(x),$$

which can be verified to be positive definite. Both $\nabla_{x_i} f_\mu(x)$ and H_i can be efficiently evaluated using network data structure for $\mathcal{G}$.

9. Numerical simulation results. In this section, we present simulation results based on the SOCP relaxations (6) and (7). Following Biswas and Ye [10, 11], we generate the true positions of the points $x_1^{\text{true}}, \ldots, x_n^{\text{true}}$ independently according to a uniform distribution on the unit square $[-.5, .5]^2$, and set $m = 0.9n$ (i.e., 10% of the points are anchors), $\mathcal{A} = \{(i, j) : \|x_i^{\text{true}} - x_j^{\text{true}}\| < \textit{radiorange}\}$, and

$$d_{ij} = \|x_i^{\text{true}} - x_j^{\text{true}}\| \cdot |1 + \epsilon_{ij} \cdot \textit{noisyfactor}| \quad \forall (i, j) \in \mathcal{A},$$

where ϵ_{ij} is a random variable representing measurement noise, and *radiorange* $\in (0, 1)$, *noisyfactor* $\in [0, 1]$. Similar to [10, 11], each ϵ_{ij} is normally distributed, and we use the parameter values of *noisyfactor* $= 0, .001, .01$ and *radiorange* $= .06$ for $n = 1000, 2000$, *radiorange* $= .035$ for $n = 4000$[7]; see Table 1.

We wrote two codes to compute an interior solution of the SOCP relaxation (6). The first code is written in MATLAB and calls SeDuMi (Version 1.05) by Jos Sturm [32], a C implementation of a predictor-corrector primal-dual interior-point method

[7] Other noise models can also be used. We use the model from [10, 11] to facilitate comparison with previous work.

TABLE 1
Input parameters for the test problems and the corresponding SOCP (7) *dimensions.* (*radiorange* $= .06$ *for* $n = 1000, 2000$, *radiorange* $= .035$ *for* $n = 4000$.)

P	n	*noisyfactor*	$\|\mathcal{A}\|$	**SOCP dim**
1	1000	0	5318	21472×28590
2	1000	.001	5068	20472×27340
3	1000	.01	5276	21304×28380
4	2000	0	21010	84440×109050
5	2000	.001	20859	83836×108295
6	2000	.01	20859	83836×108295
7	4000	0	29322	118088×154610
8	4000	.001	29322	118088×154610
9	4000	.01	29322	118088×154610

for solving SDP/SOCP, to find an interior solution of (7).[8] The second code is written in FORTRAN 77 and implements the SCGD method described in section 8, whereby we initialize $\mu = 10^{-5}$, and $x_i = x_i^{\text{true}} + \Delta_i$, with the components of Δ_i randomly generated from the square $[-.2, .2]^2$. We choose

$$\mu^{\text{final}} = 10^{-9}, \quad \psi(\mu) = \max\{10\mu, 10^{-7}\}, \quad \beta = 0.5, \quad \sigma = 0.1.$$

We choose i in step 1 in a cyclic order, and we decrease μ by a factor of 10 in step 2. These choices were made with little experimentation. Conceivably the performance can be improved with more judicious choices (e.g., replacing the cyclic order by a queue, as in the Bellman–Ford method for shortest path [5, section 2.4]).

For the interior solution $x_1, \ldots, x_m$, $(y_{ij})_{(i,j)\in\mathcal{A}}$ found, the position of the ith sensor is judged to be uniquely positioned (using Propositions 7.1 and 7.2) if there exists a $j \in N_{\mathcal{A}}(i)$ satisfying

$$\left| \|x_i - x_j\|^2 - y_{ij} \right| \le 10^{-7} d_{ij}$$

(with $x_i = x_i^{\text{true}}$ for $i > m$). In what follows, m_{up} is the number of sensors that are judged to be uniquely positioned by this test. To check the accuracy of these sensors, we compute the maximum error between their computed positions and their true positions:

$$err_{\text{up}} = \max_{i \text{ is uniquely positioned}} \|x_i - x_i^{\text{true}}\|.$$

For comparison, we also compute the maximum error between computed positions and true positions of all sensors:

$$err = \max_{i=1,\ldots,m} \|x_i - x_i^{\text{true}}\|.$$

Table 2 reports the iteration count, cpu time, the final SOCP objective value, m_{up}, err_{up}, err for the two codes. We see from Table 2 that SCGD is consistently faster than SeDuMi, though it uses more iterations. SCGD is more sensitive to *noisyfactor* than SeDuMi. We do not have a good explanation for this yet. On the other hand, the

[8] We also tried a new version 1.1 of SeDuMi, maintained by the Advanced Optimization Laboratory at McMaster University, but it gave wrong answers on our SOCP problems.

TABLE 2

Times to solve SOCP relaxation and accuracy of sensors judged to be uniquely positioned. cpu times are in minutes on an HP DL360 workstation, running MATLAB (Version 7.0) and Gnu F-77 compiler (Version 3.2.57) under Red Hat Linux 3.5.

	SeDuMi	**SCGD**
P	**iter/cpu/obj/$m_{\rm up}$/$err_{\rm up}$/err**	**iter/cpu/obj/$m_{\rm up}$/$err_{\rm up}$/err**
1	22/3.6/7.8e-6/402/7.2e-4/.11	1803189/.2/2.1e-06/357/3.8e-5/.11
2	22/3.2/9.1e-4/473/1.8e-3/.17	3523150/.4/9.1e-4/442/1.5e-3/.17
3	22/3.9/1.0e-2/554/1.5e-2/.17	14381707/1.6/1.0e-2/518/1.1e-2/.17
4	25/176.7/6.0e-6/1534/4.3e-4/.058	3482697/0.8/1.5e-5/1541/3.3e-4/.077
5	25/208.6/1.1e-2/1464/3.6e-3/.088	7894112/1.8/1.1e-2/1466/3.6e-3/.090
6	17/161.8/1.30/1710/5.1e-2/.093	12113931/2.9/1.30/1707/5.1e-2/.094
7	27/202.5/4.5e-5/2851/4.0e-4/.099	9345127/1.6/2.1e-5/2844/3.2e-4/.099
8	25/193.8/4.7e-3/2938/3.2e-3/.099	29304035/5.1/4.7e-3/2894/3.0e-3/.099
9	25/196.3/4.9e-2/3073/1.0e-2/.099	34650852/6.1/4.9e-2/3020/9.1e-3/.099

cpu times for SCGD are still high on problems with higher distance errors. These times can conceivably be further reduced by fine tuning the algorithm parameters and/or distributing the computations over multiple processors. This is a topic for future research. One idea would be to adapt the approach in [22] by terminating the SOCP method early and then applying a local descent method to the original problem (1) to refine the solution. Or we can find new methods to solve the SOCP (6), as is discussed in section 12.

We also see from Table 2 that $err_{\rm up}$ is much smaller than err and decreases with *noisyfactor*, which corroborates Propositions 7.1 and 7.2. For the larger problems 4–9, $m_{\rm up}$ is large (80–90% of m), showing that a large number of sensors are accurately positioned (with error $err_{\rm up}$) by the interior solutions found. Of course, $m_{\rm up}$ depends on *radiorange* also. If *radiorange* is small, so the graph $\mathcal{G}$ has low connectivity, then $m_{\rm up}$ would be small.

The true sensor positions and the computed positions for problems 1 and 3 are shown in Figure 7. Notice the close match of sensors whose true positions lie in the convex hull of "nearby" anchors. The positions are least accurate on the boundary, as we expect. The computed position of each sensor lies in the convex hull of its neighbors, corroborating Proposition 6.2.

At the suggestion of a referee, we also compare the SOCP solutions with solutions of the SDP (3) when n, m are small, *noisyfactor* is large, and ϵ_{ij}'s have different distributions. In particular, we apply SeDuMi to compute a solution $Z = \left[\begin{smallmatrix} Y & X^T \\ X & I_d \end{smallmatrix}\right]$ of the SDP (4), which is likely to be the analytic center solution, and extract the sensor positions $X = [x_1 \ \cdots \ x_m]$. To compare with existing work, we follow a recent study by Biswas et al. [9] of an SDP solution under noisy distance measurements and choose $n = 64$, $m = 60$, *radiorange* $= 0.3$, with 4 anchors at $(\pm.45, \pm.45)$. We also choose *noisyfactor* $\in \{0.1, 0.2\}$ and choose ϵ_{ij}'s to be (i) normally distributed, (ii) uniformly distributed on $[-\sqrt{3}, \sqrt{3}]$, or (iii) distributed as an additive-Gaussian which, with probability $\frac{1}{2}$, is normally distributed with mean 1 (otherwise with mean -1). Thus ϵ_{ij} has mean 0 and variance 1. Table 3 reports the final objective value for SDP and SOCP, as well as

$$err_{\rm rms} = \sum_{i=1}^{m} \|x_i - x_i^{\rm true}\|^2$$

and $m_{\rm up}$, $err_{\rm up}$ for SOCP.

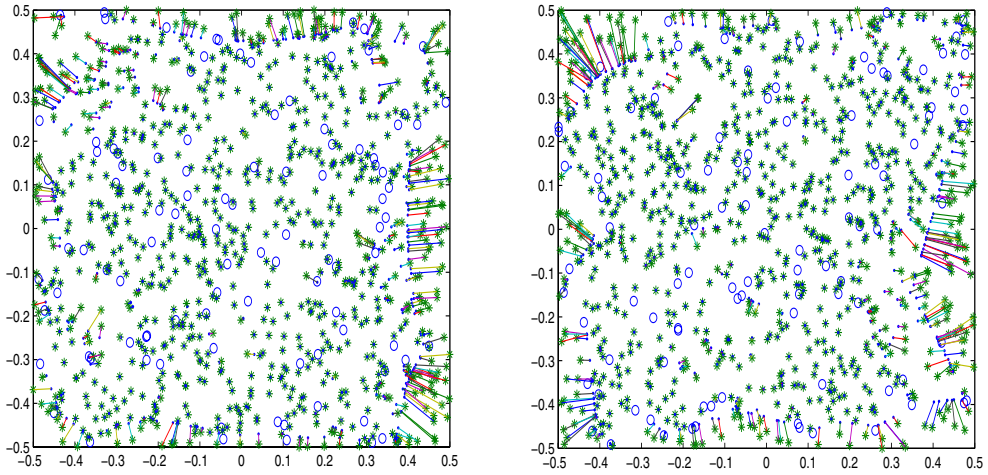


FIG. 7. *The left figure shows the anchors ("∘") and the analytic center solution found by SCGD for problem* 1 *(*$n = 1000$*). Each sensor position found ("·") is joined to its true position ("∗") by a line. The right figure shows the same information for problem* 3.

TABLE 3
Comparing analytic center solutions of SDP and SOCP for smaller problems and more noisy distance measurements.

		SOCP	**SDP**
Noise Pdf	*noisyfactor*	**obj**/$err_{\rm rms}$/$m_{\rm up}$/$err_{\rm up}$	**obj**/$err_{\rm rms}$
Normal	.1	.28/.24/52/.10	1.78/.06
	.2	.43/.48/48/.17	2.82/.23
Uniform	.1	.09/.41/30/.07	.88/.13
	.2	.24/.29/41/.11	2.26/.16
Additive-	.1	.34/.63/42/.17	2.30/.41
Gaussian	.2	.82/.71/52/.22	3.86/.50

We see from Table 3 that objective value is higher and $err_{\rm rms}$ is lower for the SDP solution than for the SOCP solution, corroborating Proposition 3.1. The $err_{\rm rms}$ is higher for both SDP and SOCP solutions under additive Gaussian noise. We do not yet have a good explanation for this. Figures 8–10 display the SDP solutions and SOCP solutions for the case of *noisyfactor* $= .2$. These results suggest that, for small randomly generated problems where the points are irregularly spaced, SDP (3) is much more preferable than SOCP (6). This situation could change with alternative problem formulations (see section 11), so further studies would be needed. In general, SOCP relaxation and mixed SDP-SOCP relaxation (see next section) seem most useful for larger problems where SDP relaxation is expensive to solve. Also, the SCGD method for solving (6) can be implemented in a highly distributed manner, with each sensor communicating with its neighbors only; see discussions at the end of section 8. This may help to reduce communication and synchronization delays among sensors in practice.

10. A mixed SDP-SOCP relaxation. Instead of an SDP or an SOCP relaxation, we can more generally consider a mixed SDP-SOCP relaxation of (1). Let $\mathcal{N}$

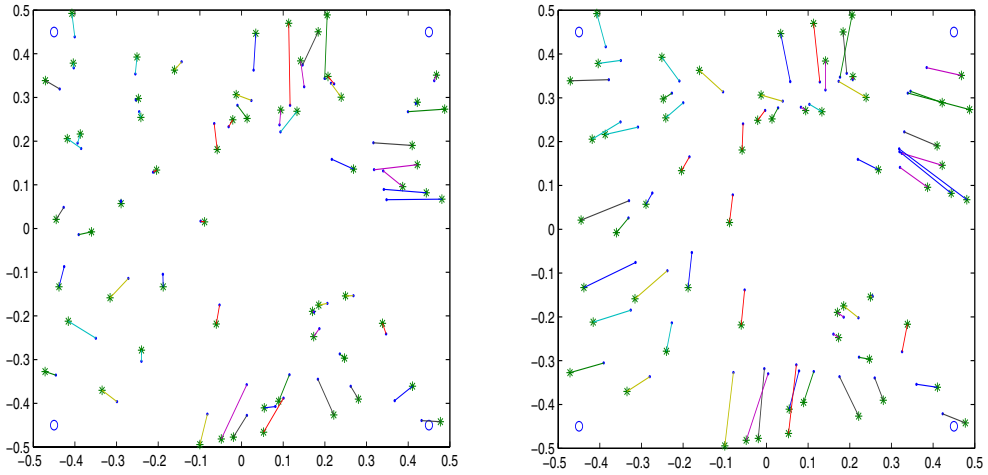


FIG. 8. *The left figure shows the anchors ("∘") and the analytic center solution of SDP found by SeDuMi for normally distributed noise and noisyfactor = .2 (row 2 of Table 3). Each sensor position found ("·") is joined to its true position ("∗") by a line. The right figure shows the same information for the analytic center solution of SOCP found by SCGD.*

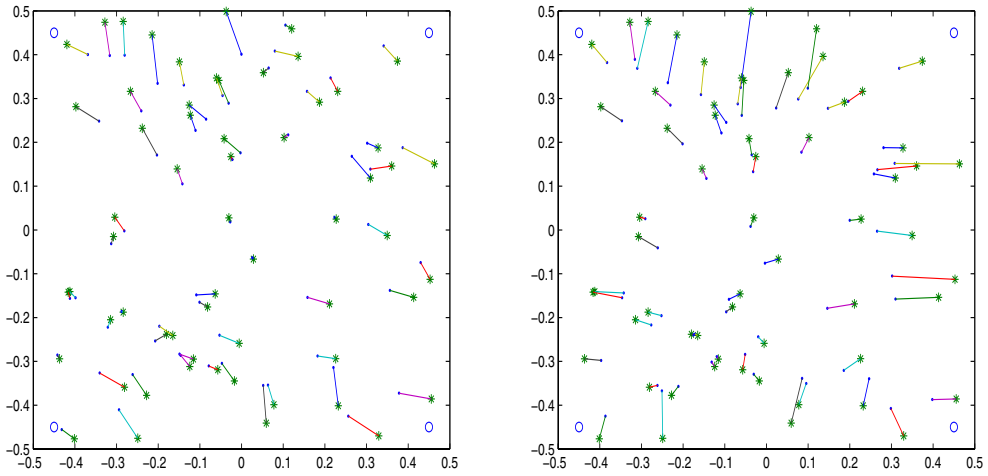


FIG. 9. *This figure is analogous to Figure 8, but for uniformly distributed noise and noisyfactor = .2 (row 4 of Table 3).*

be any subset of $\{1, \ldots, m\}$. By renumbering the points if necessary, we assume that

$$\mathcal{N} = \{\hat{m} + 1, \ldots, m\},$$

with $0 \leq \hat{m} \leq m$. Let

$$\hat{\mathcal{A}} \stackrel{\text{def}}{=} \{(i, j) \in \mathcal{A} \ : \ i \in \mathcal{N} \text{ or } j \in \mathcal{N}\}.$$

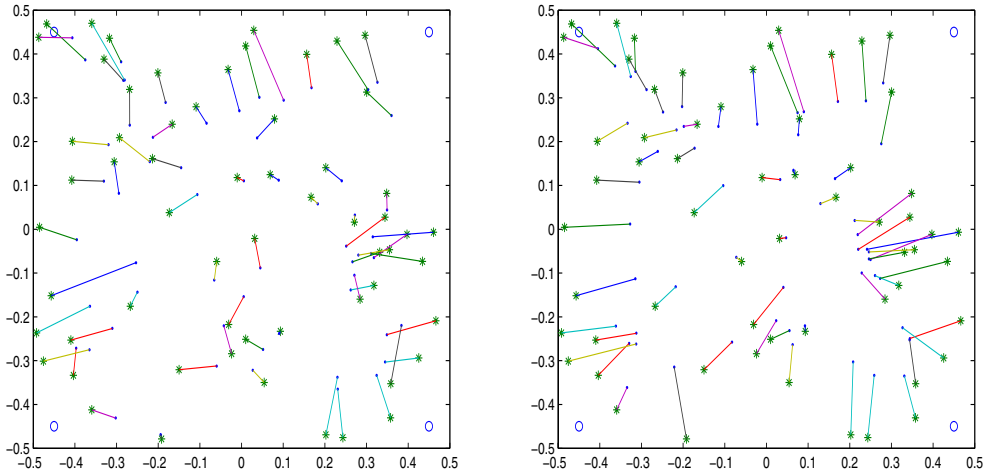


FIG. 10. *This figure is analogous to Figure* 8*, but for additive Gaussian noise and noisyfactor* = .2 *(row* 6 *of Table* 3*).*

Then the mixed SDP-SOCP relaxation associated with $\mathcal{N}$ is

$$\begin{aligned} \min_{x_1,\ldots,x_m,y_{ij},Z} \quad & \sum_{(i,j)\in\mathcal{A}\setminus\hat{\mathcal{A}}} \left|\left\langle \hat{b}_{ij}\hat{b}_{ij}^T, Z\right\rangle_F - d_{ij}^2\right| + \sum_{(i,j)\in\hat{\mathcal{A}}} |y_{ij} - d_{ij}^2| \\ \text{s.t.} \quad & [Z_{ij}]_{i,j\geq n-d} = I_d, \quad Z \succeq 0, \\ & [Z_{ij}]_{i\geq n-d, j\leq \hat{m}} = [x_1 \ \cdots \ x_{\hat{m}}], \\ & y_{ij} \geq \|x_i - x_j\|^2 \quad \forall (i,j)\in\hat{\mathcal{A}}, \end{aligned}$$

where $\hat{b}_{ij} \overset{\text{def}}{=} \left[\begin{smallmatrix} I_{\hat{m}} & 0 & 0 \\ 0 & 0 & A \end{smallmatrix}\right](e_i - e_j)$. Notice that $Z \in \mathcal{S}^{d+\hat{m}}$. This relaxation reduces to the SDP relaxation (3) if $\hat{\mathcal{A}} = \emptyset$ and reduces to the SOCP relaxation (6) if $\hat{\mathcal{A}} = \mathcal{A}$.

Such a mixed SDP-SOCP relaxation mediates between approximation accuracy and solution efficiency. In particular, Propositions 5.1 and 6.2 suggest putting into $\hat{\mathcal{A}}$ those pairs $(i,j) \in \mathcal{A}$ of sensors that are estimated to lie in the convex hull of their neighbors. Can the results in sections 4–7 be extended to the mixed SDP-SOCP relaxation? Can we design efficient methods to find interior solutions? These are topics for future research.

11. Variants of the basic problem. If "sum" is replaced by "max," then (1) becomes

$$\min_{x_1,\ldots,x_m} \quad \max_{(i,j)\in\mathcal{A}} |\|x_i - x_j\|^2 - d_{ij}^2|, \tag{31}$$

and the SDP relaxation (3) and SOCP relaxation (6) change accordingly. In general, if the objective function is a convex piecewise linear/quadratic function of $\|x_i - x_j\|^2$, $(i,j) \in \mathcal{A}$, then both an SDP relaxation and an SOCP relaxation can be analogously formulated. If the distances are not squared, then (1) becomes

$$\min_{x_1,\ldots,x_m} \sum_{(i,j)\in\mathcal{A}} |\|x_i - x_j\| - d_{ij}|. \tag{32}$$

If the distances d_{ij} are exact (i.e., $\upsilon_{\text{opt}} = 0$), then (32) is equivalent to (1). In general, (32) puts a smaller penalty on large deviation from d_{ij} and has different solutions from (1). We leave the choice of the objective function to the modeler.

For (32), an SOCP relaxation, which seems more natural than an SDP relaxation, is

$$\begin{aligned} &\min_{x_1,\ldots,x_m,y_{ij}} \sum_{(i,j)\in\mathcal{A}} |y_{ij} - d_{ij}| \\ &\text{s.t.} \quad y_{ij} \geq \|x_i - x_j\| \quad \forall (i,j) \in \mathcal{A}. \end{aligned} \tag{33}$$

By noting that $y_{ij} \geq d_{ij}$ in any solution of (33), we can write this in the standard conic form

$$\begin{aligned} \min \quad & \sum_{(i,j)\in\mathcal{A}} u_{ij} \\ \text{s.t.} \quad & x_i - x_j - w_{ij} = 0 \quad \forall (i,j) \in \mathcal{A}, \\ & y_{ij} - u_{ij} = d_{ij} \quad \forall (i,j) \in \mathcal{A}, \\ & u_{ij} \geq 0, \ (y_{ij}, w_{ij}) \in \text{Qcone}^{d+1} \quad \forall (i,j) \in \mathcal{A}, \end{aligned} \tag{34}$$

where $\text{Qcone}^{d+1} \stackrel{\text{def}}{=} \{(y,w) \in \Re \times \Re^d : y \geq \|w\|\}$ [32]. This SOCP has a smaller size than (7). In general, if the objective function is a convex piecewise linear/quadratic function of $\|x_i - x_j\|$, $(i,j) \in \mathcal{A}$, then an SOCP relaxation can be analogously formulated. Other variants of (1) involve replacing the Euclidean (ℓ_2) distance by, say, rectilinear (ℓ_1) distance or ℓ_∞ distance.

When $\upsilon_{\text{opt}} = 0$, (33) is equivalent to (6) and, moreover, they have the same analytic center solution.

12. Future directions. There are many directions for future research. For example, can our results for (1) be extended to other variants such as (31) and (32)? How do these variants compare under different distance noise distributions? What about additional constraints as discussed in [16] or replacing the 2-norm by a p-norm ($1 \leq p \leq \infty$)? Can our analysis of the SOCP relaxation (6) be extended to the mixed SDP-SOCP relaxation of section 10? Can finite termination of the SCGD method be proved? Finally, the SOCP relaxation (28) may be interpreted as the Lagrangian dual of a d-commodity convex network flow problem. For $d = 1$, this can be solved very efficiently using an ϵ-relaxation method [5, 7, 19]. Can this method be extended to $d \geq 2$, thus speeding up the solution time of the SOCP relaxation?

Acknowledgments. The author thanks Yinyu Ye for motivating the topic of this paper. He also thanks two anonymous referees for their helpful comments and suggestions.

REFERENCES

[1] A. Y. Alfakih, *Graph rigidity via Euclidean distance matrices*, Linear Algebra Appl., 310 (2000), pp. 149–165.

[2] F. Alizadeh and D. Goldfarb, *Second-order cone programming*, Math. Program., 95 (2003), pp. 3–51.

[3] J. Aspnes, D. Goldenberg, and Y. R. Yang, *On the computational complexity of sensor network localization*, in ALGOSENSORS 2004 (Turku, Finland), Lecture Notes in Comput. Sci. 3121, Springer-Verlag, New York, 2004, pp. 32–44.

[4] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*, MPS/SIAM Ser. Optim. 2, SIAM, Philadelphia, 2001.

[5] D. P. Bertsekas, *Network Optimization: Continuous and Discrete Models*, Athena Scientific, Belmont, MA, 1998.

[6] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed., Athena Scientific, Belmont, MA, 1999.

[7] D. P. Bertsekas, L. C. Polymenakos, and P. Tseng, *An ϵ-relaxation method for separable convex cost network flow problems*, SIAM J. Optim., 7 (1997), pp. 853–870.

[8] P. Biswas, T.-C. Liang, K.-C. Toh, and Y. Ye, *An SDP Based Approach for Anchor-Free 3D Graph Realization*, Report, Electrical Engineering, Stanford University, Stanford, CA, http://www.stanford.edu/~yyye/ (2005); SIAM J. Sci. Comput., submitted.

[9] P. Biswas, T.-C. Liang, K.-C. Toh, T.-C. Wang, and Y. Ye, *Semidefinite Programming Approaches for Sensor Network Localization with Noisy Distance Measurements*, Report, Electrical Engineering, Stanford University, Stanford, CA, 2005; IEEE Trans. Aut. Sci. Eng., to appear.

[10] P. Biswas and Y. Ye, *Semidefinite programming for ad hoc wireless sensor network localization*, in Proceedings of the 3rd IPSN, Berkeley, CA, 2004, pp. 46–54.

[11] P. Biswas and Y. Ye, *A distributed method for solving semidefinite programs arising from ad hoc wireless sensor network localization*, in Multiscale Optimization Methods and Applications, Nonconvex Optim. Appl. 82, Springer-Verlag, New York, 2006, pp. 69–84.

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

[13] J. M. Borwein and A. S. Lewis, *Convex Analysis and Nonlinear Optimization: Theory and Examples*, Springer-Verlag, New York, 2000.

[14] C. Chen and O. L. Mangasarian, *Smoothing methods for convex inequalities and linear complementarity problems*, Math. Programming, 71 (1995), pp. 51–69.

[15] R. Connelly, *Private communication*, Department of Mathematics, Cornell University, Ithaca, NY, 2005.

[16] L. Doherty, K. S. J. Pister, and L. El Ghaoui, *Convex position estimation in wireless sensor networks*, in Proc. 20th INFOCOM, Vol. 3, IEEE Computer Society, Los Alamitos, CA, 2001, pp. 1655–1663.

[17] T. Eren, D. K. Goldenberg, W. Whiteley, Y. R. Yang, A. S. Morse, B. D. O. Anderson, and P. N. Belhumeur, *Rigidity, computation, and randomization in network localization*, in Proc. 23rd INFOCOM, IEEE Computer Society, Los Alamitos, CA, 2004, pp. 2673–2684.

[18] M. Fukushima and L. Qi, eds., *Reformulation—Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods*, Kluwer Academic Publishers, Boston, MA, 1999.

[19] F. Guerriero and P. Tseng, *Implementation and testing of auction methods for solving separable convex cost generalized network flow problems*, J. Optim. Theory Appl., 115 (2002), pp. 113–144.

[20] S. Kim and M. Kojima, *Exact solution of some nonconvex quadratic optimization problems via SDP and SOCP relaxation*, Comput. Optim. Appl., 26 (2003), pp. 143–154.

[21] S. Kim, M. Kojima, and M. Yamashita, *Second order cone programming relaxation of a positive semidefinite constraint*, Optim. Methods Software, 18 (2003), pp. 535–541.

[22] T.-C. Liang, T.-C. Wang, and Y. Ye, *A Gradient Search Method to Round the Semidefinite Programming Relaxation Solution for Ad Hoc Wireless Sensor Network Localization*, Report, Electrical Engineering, Stanford University, Stanford, CA, http://www.stanford.edu/~yyye/ (2004).

[23] X.-D. Luo and Z.-Q. Luo, *Extensions of Hoffman's error bound to polynomial systems*, SIAM J. Optim., 4 (1994), pp. 383–392.

[24] J. J. Moré and Z. Wu, *Global continuation for distance geometry problems*, SIAM J. Optim., 7 (1997), pp. 814–836.

[25] Y. Nesterov and A. Nemirovskii, *Interior Point Polynomial Algorithms in Convex Programming*, SIAM Stud. Appl. Math. 13, SIAM, Philadelphia, 1994.

[26] J. X. Neto, O. P. Ferreira, and R. D. C. Monteiro, *Asymptotic behavior of the central path for a special class of degenerate SDP problems*, Math. Program., 103 (2005), pp. 487–514.

[27] G. Pataki, ed., *Computational Semidefinite and Second Order Cone Programming: The State of the Art*, Math. Program., 95 (2003).

[28] A. Rao, S. Ratnasamy, C. Papadimitriou, S. Shenker, and I. Stoica, *Geographic routing without location information*, in Proceedings of the 9th Annual International Conference on Mobile Computing and Networking (MobiCom '03), San Diego, CA, 2003, pp. 96–108.

[29] J. B. Saxe, *Embeddability of weighted graphs in k-space is strongly NP-hard*, in Proceedings of the 17th Allerton Conference in Communications, Control, and Computing, Monticello, IL, 1979, pp. 480–489.

[30] S. N. Simić and S. Sastry, *Distributed Localization in Wireless Ad Hoc Networks*, Report, Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, 2002; First ACM International Workshop on Wireless Sensor Networks and Applications, Atlanta, 2002, submitted.

[31] A. M.-C. So and Y. Ye, *Theory of Semidefinite Programming for Sensor Network Localization*, Report, Electrical Engineering, Stanford University, Stanford, CA, 2004; in SODA 2005; Math. Program., to appear.

[32] J. F. Sturm, *Using Sedumi* 1.02*, A MATLAB* Toolbox for Optimization over Symmetric Cones (Updated for Version* 1.05*)*, Report, Department of Econometrics, Tilburg University, Tilburg, The Netherlands, 1998–2001.

[33] T. Wang and J.-S. Pang, *Global error bounds for convex quadratic inequality systems*, Optimization, 31 (1994), pp. 1–12.

SIAM J. OPTIM.
Vol. 18, No. 1, pp. 186–205

APPROXIMATING K-MEANS-TYPE CLUSTERING VIA SEMIDEFINITE PROGRAMMING*

JIMING PENG† AND YU WEI‡

Abstract. One of the fundamental clustering problems is to assign n points into k clusters based on minimal sum-of-squared distances (MSSC), which is known to be NP-hard. In this paper, by using matrix arguments, we first model MSSC as a so-called 0-1 semidefinite programming (SDP) problem. We show that our 0-1 SDP model provides a unified framework for several clustering approaches such as normalized k-cut and spectral clustering. Moreover, the 0-1 SDP model allows us to solve the underlying problem approximately via the linear programming and SDP relaxations. Second, we consider the issue of how to extract a feasible solution of the original 0-1 SDP model from the optimal solution of the relaxed SDP problem. By using principal component analysis, we develop a rounding procedure to construct a feasible partitioning from a solution of the relaxed problem. In our rounding procedure, we need to solve a K-means clustering problem in $\Re^{k-1}$, which can be done in $O(n^{k^2-2k+2})$ time. In case of biclustering, the running time of our rounding procedure can be reduced to $O(n \log n)$. We show that our algorithm provides a 2-approximate solution to the original problem. Promising numerical results for biclustering based on our new method are reported.

Key words. K-means clustering, principal component analysis, 0-1 SDP, relaxation, computational complexity, approximation

AMS subject classifications. 90C22, 68T10

DOI. 10.1137/050641983

1. Introduction.

In general, clustering involves partitioning a given data set into subsets based on the closeness or similarity among the data. Clustering is one of the major issues in data mining and machine learning with many applications arising from different disciplines including text retrieval, pattern recognition, and web mining [19, 23].

There are many kinds of clustering problems and algorithms, resulting from various choices of measurements used in the model to measure the similarity/dissimilarity among entities in a data set. For a comprehensive introduction to the topic, we refer the reader to the books [19, 23], and for more recent results, see survey papers [9] and [20].

Among various criteria in clustering, the minimum sum-of-squared Euclidean distance (MSSC) from each entity to its assigned cluster center is the most intuitive and broadly used. In the present paper, we are particularly interested in the partitioning procedure for MSSC. A well-known method to deal with this problem is the classical K-means [28]. To describe the algorithm, let us go into a bit more detail.

*Received by the editors October 5, 2005; accepted for publication (in revised form) September 11, 2006; published electronically February 26, 2007. The main part of this work was done when the second author was a master student in the Advanced Optimization Lab, Department of Computing and Software, McMaster University under the supervision of the first author. The research is supported by the grant RPG 249635-02 of the National Sciences and Engineering Research Council of Canada (NSERC) and a PREA award. This research is also supported by the MITACS project "New Interior Point Methods and Software for Convex Conic-Linear Optimization and Their Application to Solve VLSI Circuit Layout Problems."

http://www.siam.org/journals/siopt/18-1/64198.html

†Corresponding author. Advanced Optimization Lab, Department of Computing and Software, McMaster University, Hamilton, ON L8S 4K1, Canada (pengj@mcmaster.ca).

‡Advanced Optimization Lab, Department of Computing and Software, McMaster University, Hamilton, ON L8S 4K1, Canada (weiy3@mcmaster.ca).

Given a set S of n points in a d-dimensional Euclidean space,[1] denoted by

$$S = \{\mathbf{s}_i = (s_{i1}, \ldots, s_{id})^T \in \mathbf{R}^d, \quad i = 1, \ldots, n\},$$

the task of a partitional MSSC is to find an assignment of the n points into k disjoint clusters $\mathcal{S} = (S_1, \ldots, S_k)$ centered at cluster centers $\mathbf{c}_j\,(j = 1, \ldots, k)$ such that the total sum-of-squared Euclidean distances from each point $\mathbf{s}_i$ to its assigned cluster centroid $\mathbf{c}_i$

$$f(S, \mathcal{S}) = \sum_{j=1}^{k} \sum_{i=1}^{|S_j|} \left\| \mathbf{s}_i^{(j)} - \mathbf{c}_j \right\|^2$$

is minimized. Here $|S_j|$ is the number of points in S_j, and $\mathbf{s}_i^{(j)}$ is the ith point in S_j. Note that if the cluster centers are known, then the function $f(S, \mathcal{S})$ achieves its minimum when each point is assigned to its closest cluster center. Therefore, MSSC can be described by the following bilevel programming problem (see, for instance, [4, 30]):

$$\min_{\mathbf{c}_1, \ldots, \mathbf{c}_k} \sum_{i=1}^{n} \min\{\|\mathbf{s}_i - \mathbf{c}_1\|^2, \ldots, \|\mathbf{s}_i - \mathbf{c}_k\|^2\}. \tag{1}$$

Geometrically speaking, assigning each point to the nearest center fits into a framework called the *Voronoi program*, and the resulting partition is named the *Voronoi partition*. On the other hand, if the points in cluster S_j are fixed, then the function

$$f(S_j, \mathcal{S}_j) = \sum_{i=1}^{|S_j|} \left\| \mathbf{s}_i^{(j)} - \mathbf{c}_j \right\|^2$$

is minimized when

$$\mathbf{c}_j = \frac{1}{|S_j|} \sum_{i=1}^{|S_j|} \mathbf{s}_i^{(j)}.$$

The classical K-means algorithm [28], based on the above two observations, is described as follows.

K-means clustering algorithm.

(1) Randomly generate k cluster centers in a domain containing all the points,
(2) Assign each point to the closest cluster center,
(3) Recompute the cluster centers using the current cluster memberships,
(4) If a convergence criterion is met, stop; Otherwise go to step 2.

In spite of its popularity, the above simple procedure is very sensitive to the initial choice of the starting points and could not find the global solution in terms of its objective in general.

Another way to model MSSC works as follows. Let $X = [x_{ij}] \in \Re^{n \times k}$ be the assignment matrix defined by

$$x_{ij} = \begin{cases} 1 & \text{if } \mathbf{s}_i \text{ is assigned to } S_j; \\ 0 & \text{otherwise.} \end{cases}$$

[1] In the present paper, we always assume that $n \geq k > 1$, because otherwise the underlying clustering problem becomes trivial.

As a consequence, the cluster center of the cluster S_j, as the mean of all the points in the cluster, is defined by

$$\mathbf{c}_j = \frac{\sum_{l=1}^{n} x_{lj}\mathbf{s}_l}{\sum_{l=1}^{n} x_{lj}}.$$

Using this fact, we can represent (1) as

$$\min_{x_{ij}} \sum_{j=1}^{k}\sum_{i=1}^{n} x_{ij} \left\| \mathbf{s}_i - \frac{\sum_{l=1}^{n} x_{lj}\mathbf{s}_l}{\sum_{l=1}^{n} x_{lj}} \right\|^2 \tag{2}$$

$$\text{s.t.} \sum_{j=1}^{k} x_{ij} = 1 \ (i = 1, \ldots, n), \tag{3}$$

$$\sum_{i=1}^{n} x_{ij} \geq 1 \ (j = 1, \ldots, k), \tag{4}$$

$$x_{ij} \in \{0, 1\} \ (i = 1, \ldots, n;\ j = 1, \cdots, k). \tag{5}$$

The constraint (3) ensures that each point $\mathbf{s}_i$ is assigned to one and only one cluster, and constraint (4) ensures that there are exactly k clusters. This is a mixed integer programming problem with a nonlinear objective function [14], which is NP-hard. The problem has two difficulties. First, the constraints are discrete. Second, the objective is nonlinear and nonconvex. Both difficulties make MSSC extremely hard to solve.

Many different approaches have been proposed for attacking (2) both in the communities of machine learning and optimization [1, 14, 8]. Most methods are heuristics that can locate only a good local solution and not the exact global solution for (2). Only a few works are dedicated to the exact algorithm for (2) as listed in the references of [8]. In particular, by using the notion of *Voronoi partitions*, Inaba, Katoh, and Imai [18] showed that the exact solution of (2) can be located in $O(n^{kd+1})$ time. Due to its high complexity, the algorithm can be applied only to small data sets. In the special case $k = 2$, Hansen, Jaumard, and Mladenović [15] provided an algorithm running in $O(n^{d+1} \log n)$ time, where d is the dimension of the space to which the entities belong. Promising numerical results are reported for data sets in low dimensions.

Approximation methods provide a useful approach for solving (2). There are several ways to approximate (2). For example, Hasegawa et al. [16] proposed to select the k points from the present data set and then run the classical K-means for the selected centers. If we try all possible combinations of the k starting centers and output the best solution as the final one, then we can obtain a 2-approximately optimal solution for (2) in $O(n^{k+1})$ time [16]. In [33], Mutousek proposed a geometric approximation method that can find a $(1 + \epsilon)$ approximately optimal solution for (2) in $O(dn \log^k n)$ time, where the constant hidden in the big-O notation depends polynomially on ϵ^{-1}. Though theoretically efficient, so far no numerical results have been reported based on Matousek's algorithm. More recently, Kumar, Sabharwal, and Sen [25] proposed a linear time $O(2^{\epsilon^{-O(1)}} nd)$ algorithm for K-means clustering based on random sampling techniques and showed that their algorithm can find a $(1 + \epsilon)$ approximation with a certain probability. Only theoretical analysis based on the probabilistic model is presented, however. For more results on approximation algorithms for K-means clustering based on randomization techniques, we refer the reader to [25] and the references therein.

An efficient way of approximation is to attack the original problem (typically NP-hard) by solving a relaxed polynomially solvable problem. This has been well studied in the field of optimization, in particular, in the areas of combinatorial optimization and semidefinite programming (SDP) [10]. We noted that recently, Xing and Jordan [44] considered the SDP relaxation for the so-called normalized k-cut spectral clustering.

In the present paper, we mainly focus on developing approximation methods for (2) based on SDP relaxation. A crucial step in relaxing (2) is to rewrite the objective in (2) as a simple convex function of matrix argument that can be tackled easily, while the constraint set still enjoys certain geometric properties. This idea was possibly first suggested in [12], where the authors owed the idea to an anonymous referee. However, the authors of [12] did not explore the idea in depth to design any usable algorithm. A similar effort was made in [45], where the authors rewrote the objective in (2) as a convex quadratic function in which the argument is an $n \times k$ orthonormal matrix.

One major contribution of the present paper is the introduction of a new optimization model (0-1 SDP), which follows the same stream as in [12, 45]. However, different from the approach in [45], where the authors used only a quadratic objective and simple spectral relaxation, we elaborate more on how to characterize (2) exactly by means of matrix arguments. In particular, we show that MSSC can be modeled as the so-called 0-1 SDP, which can be further relaxed to polynomially solvable linear programming (LP) and SDP. Our model provides novel avenues not only for solving MSSC, but also for solving clustering problems based on some other criteria. For example, the K-means clustering in the kernel space and the clustering problem based on normalized cuts can also be embedded into our model. Moreover, by slightly changing the constraints in the 0-1 SDP model, we can attack clustering problems with constraints, e.g., the so-called balanced clustering.

The second major contribution of the present work is the development of an efficient approximation algorithm for the 0-1 SDP model, especially for biclustering. For this purpose, we first relax the 0-1 SDP model by removing some constraints so that the relaxed problem can be solved by using singular value decomposition (SVD) of the underlying coefficient matrix. Then we introduce a rounding procedure to extract a feasible solution for the original 0-1 model. In our rounding procedure, we need to solve the K-means clustering problem in $\Re^{k-1}$, which can be done in $O(n^{k^2-2k+2})$ time. We show that our algorithm provides a 2-approximate solution to the original K-means clustering.

Our algorithm uses an idea similar to the so-called spectral clustering [7, 45, 34], where the SVD of the coefficient matrix is employed to calculate the eigenvectors corresponding to the first k largest eigenvalues of the coefficient matrix, and these eigenvectors are further used to formulate a new data set in a lower dimension for which the classical K-means clustering is performed. A similar idea has been adopted in the principal components analysis (PCA) [22, 5]. In particular, Drineas et al. [7] proposed to solve the K-means clustering problem in $\Re^k$ whose solution can be found in $O(n^{k^2+1})$ time and showed that their method can provide a 2-approximate solution to the original K-means clustering problem.[2] For the classical K-means clustering, our algorithm can be viewed as a slight improvement over the algorithm in [7]. As we

[2]It should be pointed out that although the algorithm in [7] enjoys some nice properties, the high complexity in solving the subproblem might prevent it from practical efficiency when dealing with large-scale data sets. For example, for the biclustering problem with a data set with $n = 10^4$, the running time of the algorithm in [7] will reach a formidable $O(10^{20})$.

shall see later in our discussion, the only difference between the two algorithms in [7] and that in the present paper lies in how the 0-1 SDP model is relaxed.

It should be mentioned that for the cases with $k \geq 3$, the complexity of our algorithm is not as good as that of the algorithms in [16, 33]. However, in the case of biclustering ($k = 2$, which is still NP-hard), the global solution to the subproblem in our algorithm can be found in $O(n \log n)$ time. This implies that, for the classical K-means clustering problems with $k = 2$ and $d << n$, the complexity of our algorithm reduces to $O(n \log n)$. Although the theoretical complexity of our algorithm is only slightly better than that of the algorithm in [33], our algorithm is practically much simpler and allows us to efficiently solve large-scale biclustering problems in many applications with guaranteed quality.

Besides the above-mentioned contributions, our algorithm also provides a useful tool for solving several new clustering problems such as the balanced K-means clustering for which no approximation algorithms have been reported in the literature. The algorithm is particularly helpful for solving clustering problems arising from text mining where the data is typically in very high dimension ($n << d$) since the original data is only used to calculate the coefficient matrix.

The paper is organized as follows. In section 2, we show that MSSC can be modelled as a 0-1 SDP, which allows convex relaxation such as SDP and LP. In section 3, we consider approximation algorithms for solving our 0-1 SDP model. We propose to use PCA to reduce the dimension of the problem, and then perform K-means clustering in the lower dimension. The approximate ratio between the obtained solution and the global solution of the original K-means clustering is estimated in section 4. In section 5, we report some preliminary numerical tests, and finally we close the paper by a few concluding remarks.

2. 0-1 SDP for clustering problems. In this section, we establish the equivalence between several clustering scenarios and the 0-1 SDP model. The section has three parts. In the first part, we briefly describe SDP and 0-1 SDP. In the second part, we establish the equivalence between MSSC and the 0-1 SDP model. In the last part, we explore the interrelation between the 0-1 SDP model and other K-means-type clustering problems.

2.1. 0-1 semidefinite programming. In general, SDP refers to the problem of minimizing (or maximizing) a linear function over the intersection of a polyhedron and the cone of symmetric and positive semidefinite matrices. The canonical SDP takes the following form:

$$(\text{SDP}) \begin{cases} \min & \text{Tr}(WZ) \\ \text{s.t.} & \text{Tr}(B_i Z) = b_i \quad \text{for } i = 1, \ldots, m, \\ & Z \succeq 0. \end{cases}$$

Here $\text{Tr}(M)$ denotes the trace of the matrix M, and $Z \succeq 0$ means that Z is positive semidefinite. If we replace the constraint $Z \succeq 0$ by the requirement that $Z^2 = Z$ and $Z = Z^T$, then we end up with the following problem:

$$(\text{0-1 SDP}) \begin{cases} \min & \text{Tr}(WZ) \\ \text{s.t.} & \text{Tr}(B_i Z) = b_i \quad \text{for } i = 1, \ldots, m, \\ & Z^2 = Z, Z = Z^T. \end{cases}$$

We call it 0-1 SDP owing to the similarity of the constraint $Z^2 = Z$ to the classical 0-1 requirement in integer programming.

2.2. Equivalence of MSSC to 0-1 SDP. In this subsection we give the 0-1 SDP model of MSSC, which was first established in [36]. However, for completeness, we still give a detailed description of the reformulation process.

By rearranging the items in the objective of (2), we have

$$f(S,\mathcal{S}) = \sum_{i=1}^{n} \|\mathbf{s}_i\|^2 \left(\sum_{j=1}^{k} x_{ij}\right) - \sum_{j=1}^{k} \frac{\left\|\sum_{i=1}^{n} x_{ij}\mathbf{s}_i\right\|^2}{\sum_{i=1}^{n} x_{ij}} = \mathrm{Tr}\big(W_S W_S^T\big) - \sum_{j=1}^{k} \frac{\left\|\sum_{i=1}^{n} x_{ij}\mathbf{s}_i\right\|^2}{\sum_{i=1}^{n} x_{ij}}, \tag{6}$$

where $W_S \in \Re^{n\times d}$ denotes the matrix whose ith row is $\mathbf{s}_i^T$. Since X is an assignment matrix, we have

$$X^T X = \mathrm{diag}\left(\sum_{i=1}^{n} x_{i1}^2, \ldots, \sum_{i=1}^{n} x_{ik}^2\right) = \mathrm{diag}\left(\sum_{i=1}^{n} x_{i1}, \ldots, \sum_{i=1}^{n} x_{ik}\right).$$

Let

$$Z := [z_{ij}] = X(X^T X)^{-1} X^T;$$

we can write (6) as $\mathrm{Tr}\big(W_S W_S^T (I - Z)\big) = \mathrm{Tr}\big(W_S^T W_S\big) - \mathrm{Tr}\big(W_S^T W_S Z\big)$. Obviously Z is a projection matrix satisfying $Z^2 = Z$ with nonnegative elements. For any integer m, let e^m be the vector in $\Re^m$ with all entries equal to 1. We can write the constraint (3) as

$$Xe^k = e^n.$$

It follows immediately that

$$Ze^n = ZXe^k = Xe^k = e^n.$$

Moreover, the trace of Z should equal to k, the number of clusters, i.e.,

$$\mathrm{Tr}(Z) = k.$$

Therefore, we have the following 0-1 SDP model for MSSC:

$$\begin{aligned} \min & \mathrm{Tr}\big(W_S W_S^T (I - Z)\big) \\ & Ze = e, \mathrm{Tr}(Z) = k, \\ & Z \geq 0, Z = Z^T, Z^2 = Z. \end{aligned} \tag{7}$$

Here $Z \geq 0$ means the componentwise inequality, and the constraints $Z^T = Z$ and $Z^2 = Z$ imply that Z is an orthogonal projection matrix.

The following result from [36] established the equivalence between the 0-1 SDP model (7) and MSSC.

THEOREM 2.1. *Solving the* 0–1 *SDP problem* (7) *is equivalent to finding a global solution of the integer programming problem* (2).

It is worthwhile comparing the two objective functions in (7) and (2). First, the objective function in (7) is linear while the constraint in (7) is still nonlinear. The most difficult part in the constraint of (7) is the requirement that $Z^2 = Z$. Several relaxations of problem (7) will be discussed in the next section.

2.3. 0-1 SDP reformulation for other clustering approaches. In this subsection, we show that the 0-1 SDP can also be used for other clustering approaches based on other measurements. Let us consider the more general 0-1 SDP model for clustering,

$$\begin{aligned} \min & \mathrm{Tr}(W(I-Z)) \\ & Ze = e, \mathrm{Tr}(Z) = k, \\ & Z \geq 0, Z^2 = Z, Z = Z^T, \end{aligned} \tag{8}$$

where W is the so-called affinity matrix whose entries represent the similarities or closeness among the entities in the data set. In the MSSC model, we use the geometric distance between two points to characterize the similarity between them. In this case, we have $W_{ij} = \mathbf{s}_i^T \mathbf{s}_j$. However, we can also use a general function $\phi(\mathbf{s}_i, \mathbf{s}_j)$ to describe the similarity relationship between $\mathbf{s}_i$ and $\mathbf{s}_j$. For example, let us choose

$$W_{ij} = \phi(\mathbf{s}_i, \mathbf{s}_j) = \exp^{-\frac{\|\mathbf{s}_i - \mathbf{s}_j\|^2}{\sigma}}, \quad \sigma > 0. \tag{9}$$

This leads to the so-called K-means clustering in the kernel space. In order to apply the classical K-means algorithm to (8), we can first use the singular eigenvalue decomposition method to decompose the matrix W into the product of two matrices, i.e., $W = U^T U$. In this case, each column of U can be cast as a point in a suitable space. Then, we can apply the classical K-means method for the MSSC model to solving problem (8). This is exactly the procedure that the recently proposed spectral clustering follows [2, 34, 43, 44]. However, we now have a new interpretation for spectral clustering, i.e., a variant of MSSC in a different kernel space. It is worthwhile mentioning that certain variants of K-means can be adapted to solve (8) directly without using the SVD of the affinity matrix.

We note that recently, the normalized cut and spectral clustering have attracted a lot of attention in the machine learning community, and many interesting results about these two approaches have been reported [13, 31, 34, 40, 43, 44, 45]. In particular, Zha et al. [45] discussed the links between spectral relaxation and K-means clustering. Similar ideas were also used in [34]. An SDP relaxation for the normalized k-cut problem was discussed in [44]. For completeness, we next describe briefly how the normalized cut problem can be embedded into the 0-1 SDP model. Let us first recall the exact model for the normalized k-cut problem [44]. Let W be the affinity matrix defined by (9) and X be the assignment matrix in the set $\mathcal{F}_k$ defined by

$$\mathcal{F}_k = \{X : Xe^k = e^n, x_{ij} \in \{0, 1\}\}.$$

Let $d = We^n$ and $D = \mathrm{diag}\,(d)$. The exact model for the normalized k-cut problem in [44] can be rewritten as

$$\max_{X \in \mathcal{F}_k} \mathrm{Tr}\big((X^T D X)^{-1} X^T W X\big). \tag{10}$$

If we define

$$Z = D^{\frac{1}{2}} X (X^T D X)^{-1} X^T D^{\frac{1}{2}},$$

then we have

$$Z^2 = Z, Z^T = Z, Z \geq 0, Z d^{\frac{1}{2}} = d^{\frac{1}{2}}.$$

Following a similar process as in the proof of Theorem 2.1, we can show that the model (10) is equivalent to the following 0-1 SDP:

$$\begin{aligned} \min & \mathrm{Tr}\Big(D^{-\frac{1}{2}} W D^{-\frac{1}{2}} (I - Z)\Big) \\ & Z d^{\frac{1}{2}} = d^{\frac{1}{2}}, \mathrm{Tr}(Z) = k, \\ & Z \geq 0, Z^2 = Z, Z = Z^T. \end{aligned} \tag{11}$$

The only difference between (8) and (11) is the introduction of the scaling matrix D.

Besides the above-mentioned cases, the 0-1 SDP model can also be applied to the so-called balanced clustering [3], where the number of entities in every cluster is restricted. One special case of balanced clustering is requiring the number of entities in every cluster to be equal to or larger than a prescribed quantity, i.e., $|C_i| \geq \tilde{n}$. It is straightforward to see that such a problem can be modeled as a 0-1 SDP by adding the constraint $Z_{ii} \leq \frac{1}{\tilde{n}}$ to (8), which leads to the following problem:

$$\begin{aligned} \min & \mathrm{Tr}(W(I - Z)) \\ & Z_{ii} \leq \frac{1}{\tilde{n}}, \quad i = 1, \ldots, n, \\ & Ze = e, \mathrm{Tr}(Z) = k, \\ & Z \geq 0, Z^2 = Z, Z = Z^T. \end{aligned} \tag{12}$$

3. Approximation algorithms for solving 0-1 SDP. In this section we discuss how to solve the 0-1 SDP model for clustering. For simplification of our discussion, we restrict ourselves to the model (8) with a positive semidefinite matrix W. This assumption is satisfied in the MSSC model as well as in the so-called kernel K-means clustering, where the kernel matrix is defined by (9) or some other kernel function. It is worthwhile mentioning that although we restrict our discussion to (8), with minimal effort our results can be extended to (11) as well.

The section consists of two parts. In the first subsection, we give a general introduction to algorithms for solving (8). In the second part, we introduce a new approximation method for (8).

3.1. Algorithms for 0-1 SDP. In this subsection, we discuss various algorithms for solving the 0-1 SDP model (8). From an algorithm design viewpoint, we can categorize all the algorithms for (8) into two groups. The first group consists of the so-called feasible iterative algorithms, where all the generated iterates are feasible for problem (8) and the objective function value decreases step by step until some termination criterion is reached. The classical K-means algorithm described in the introduction can be interpreted as a special feasible iterative scheme for attacking (8). It is also easy to see that many variants of the K-means algorithm, such as the variants proposed in [17, 21], can also be interpreted as specific iterative schemes for (8).

The second group of algorithms for (8) consists of approximation algorithms that are based on LP/SDP relaxation. We start with a general procedure for those algorithms.

Approximation Algorithm Based on Relaxation.

Step 1. Choose a relaxation model for (7),

Step 2. Solve the relaxed problem for an approximate solution,

Step 3. Use a rounding procedure to extract a feasible solution to (7) from the approximate solution.

Various relaxations and rounding procedures have been proposed for solving (8) in the literature. For example, in [36], Peng and Xia considered a relaxation of (8) based on linear programming and a rounding procedure was also proposed in that work. Xing and Jordan [44] considered an SDP relaxation for the normalized cut problem and proposed a rounding procedure based on the SVD of the solution Z of the relaxed problem, i.e., $Z = U^T U$. In their approach, every row of U^T is cast as a point in the new space, and then the weighted K-means clustering is performed over the new set of points in $\Re^k$. Similar works for spectral clustering can also be found in [13, 31, 34, 43, 45], where the SVD of the underlying matrix W is used and a K-means-type clustering based on the eigenvectors of W is performed. In the above-mentioned works, the solutions obtained from the weighted K-means algorithm for the original problem and that based on the eigenvectors of W has been compared, and simple theoretical bounds have been derived based on the eigenvalues of W.

The idea of using the SVD of the underlying matrix W is natural in the so-called PCA [22]. In [5], the link between PCA and K-means clustering was also explored and simple bounds were derived. In particular, Drineas et al. [7] proposed using PCA to reduce the dimension of the input data, and then performing K-means clustering in the reduced subspace $\Re^k$. They proved that their algorithm can provide a 2-approximation solution to the clustering problem in the original input space.

We note that in [40], Shi and Malik used the eigenvector of a projection matrix of W (not W itself) onto a subspace to construct a feasible partitioning for the original problem. In this paper, we follow a similar idea as that in [40]. We first use SVD to obtain the $k-1$ eigenvectors corresponding to the first $k-1$ largest eigenvalues of a projection matrix of W, and then perform K-means clustering in $\Re^{k-1}$. This allows us to improve the complexity of the algorithm for solving the subproblem in the reduced space. As we shall see later, such a rounding procedure can also provide a 2-approximation solution to the original problem.

3.2. A new approximation method. In this subsection, we describe our SDP-based approximation method for (8). We start our discussion on various relaxation forms for (8).

First, recall that in (8), the argument Z is stipulated to be a projection matrix, i.e., $Z^2 = Z$ and $Z = Z^T$. This implies that the matrix Z is a positive semidefinite matrix whose eigenvalues are either 0 or 1. A straightforward relaxation to (8) is replacing the requirement $Z^2 = Z$ by the relaxed condition

$$I \succeq Z \succeq 0.$$

Note that in (8), we further stipulate that all the entries of Z are nonnegative, and the sum of each row (or each column) of Z equals to 1. This means all the eigenvalues of Z are less than or equal to 1. In this circumstance, the constraint $Z \preceq I$ becomes superfluous and can be removed. Therefore, we obtain the following SDP relaxation[3]:

$$\begin{aligned} \min & \mathrm{Tr}(W(I-Z)) \\ & Ze = e, \mathrm{Tr}(Z) = k, \\ & Z \geq 0, Z \succeq 0. \end{aligned} \tag{13}$$

The above problem is feasible and bounded below. We can apply many existing optimization solvers such as interior-point methods to solve (13). It is known that an

[3]In [44], the constraint $Ze = e$ in (8) is replaced by $Zd = d$, where d is a positive scaling vector associated with the affinity matrix.

approximate solution to (13) can be found in polynomial time. However, we would like to point out here that although there exist theoretically polynomial algorithms for solving (13), most of the present optimization solvers are unable to handle the problem in large size efficiently.

Another interesting relaxation to (8) is to further relax (13) by dropping some constraints. For example, if we remove the nonnegative requirement on the elements of Z, then we obtain the following simple SDP problem[4]:

$$\begin{aligned} \min & \mathrm{Tr}(W(I-Z)) \\ & Ze = e, \mathrm{Tr}(Z) = k, \\ & I \succeq Z \succeq 0. \end{aligned} \tag{14}$$

In what follows we discuss how to solve (14). Note that if Z is a feasible solution for (14), then we have

$$\frac{1}{\sqrt{n}} Ze = \frac{1}{\sqrt{n}} e,$$

which implies $\frac{1}{\sqrt{n}} e$ is an eigenvector of Z corresponding to its largest eigenvalue 1. For any feasible solution of (14), let us define

$$Z_1 = Z - \frac{1}{n} ee^T.$$

It is easy to see that

$$Z_1 = \left(I - \frac{1}{n} ee^T\right) Z = \left(I - \frac{1}{n} ee^T\right) Z \left(I - \frac{1}{n} ee^T\right); \tag{15}$$

i.e., Z_1 represents the projection of the matrix Z onto the null subspace of e. Moreover, it is easy to verify that

$$\mathrm{Tr}(Z_1) = \mathrm{Tr}(Z) - 1 = k - 1.$$

Let W_1 denote the projection of the matrix W onto the null space of e, i.e.,

$$W_1 = \left(I - \frac{1}{n} ee^T\right) W \left(I - \frac{1}{n} ee^T\right). \tag{16}$$

Then, we can reduce (14) to

$$\begin{aligned} \min & \mathrm{Tr}(W_1(I - Z_1)) \\ & \mathrm{Tr}(Z_1) = k - 1, \\ & I \succeq Z_1 \succeq 0. \end{aligned} \tag{17}$$

Let $\lambda_1, \ldots, \lambda_{n-1}$ be the eigenvalues of the matrix W_1 listed in the order of decreasing values. The optimal solution of (17) can be achieved if and only if [35]

$$\mathrm{Tr}(W_1 Z_1) = \sum_{i=1}^{k-1} \lambda_i.$$

[4]We point out that in [7], the authors considered a relaxation of the K-means clustering based on projection matrices of rank k. This relaxation is equivalent to simply removing the nonnegativity requirement on the elements of Z and the constraint $Ze = e$. In such a case, solving the relaxed problem in [7] reduces to the standard PCA.

This gives us an easy way to solve (17) and correspondingly (14). The algorithmic scheme for solving (14) can be described as follows.

Relaxation Algorithm 1.

Step 1. Calculate the projection W_1 via (16);

Step 2. Use the SVD method to compute the first $k-1$ largest eigenvalues of the matrix W_1 and their corresponding eigenvectors $v^1, \ldots, v^{k-1}$,

Step 3. Set

$$Z = \frac{1}{n} ee^T + \sum_{i=1}^{k-1} v^i v^{i^T}.$$

From our above discussion, we immediately have the following theorem.

THEOREM 3.1. *Let Z^* be the global optimal solution of* (8)*, and $\lambda_1, \ldots, \lambda_{k-1}$ be the first largest eigenvalues of the matrix W_1. Then we have*

$$\mathrm{Tr}(W(I - Z^*)) \geq \mathrm{Tr}(W) - \frac{1}{n} e^T W e - \sum_{i=1}^{k-1} \lambda_i.$$

It is of interest to discuss briefly the computational cost of the relaxation algorithm 1. In general, to compute the projection matrix W_1, we need $O(n^2)$ operations in total. Typically, performing the SVD for W_1 requires $O(n^3)$ time. If we use the power method or some other iterative methods [11] to compute the eigenvectors corresponding to the first $k-1$ largest eigenvalues of W_1, then the overall computational cost can be reduced to $O(kn^2)$. It should be mentioned that for the classical K-means clustering, we do not even need to compute the matrices W and W_1 explicitly. Recall that in case of the classical K-means clustering, we have $W = W_s W_S^T$. It is straightforward to see that for calculating W_1, we can first perform a simple normalization for the data set $s_i = s_i - \bar{s}$, where $\bar{s}$ is the geometric center of the whole data set. Correspondingly, the task in Step 2 of the relaxation algorithm can be realized by performing the SVD on the new coefficient matrix $W_{\bar{S}}$ for the normalized data set, which can be done in $O(\min\{n,d\}^3 + \min\{n,d\}nd)$ time.[5] If $d << n$ (this is true for data sets in many applications), the computational cost involved in the relaxation algorithm is linear in n.

We point out that if $k = 2$, then Step 2 in the above algorithm uses the eigenvector corresponding to the largest eigenvalue of W_1. A similar relaxation was used by Shi and Malik [40] (see also [43]) for image segmentation where the normalized cut problem with $k = 2$ was discussed. Similar bounds for the normalized cut problem and spectral clustering can also be found in [34, 5].

Note that solving the relaxed problem (14) cannot provide a solution to the original problem (8). In what follows we propose a rounding procedure to extract a feasible solution for (8) from a solution of the relaxed problem (14). Our rounding procedure follows a similar vein as the rounding procedure in [7]. Let us denote $V = (\sqrt{\lambda_1} v^1, \ldots, \sqrt{\lambda_{k-1}} v^{k-1}) \in \Re^{n \times (k-1)}$ the solution matrix obtained from the relaxation algorithm, Algorithm 1. We can cast each row of V as a point in $\Re^{k-1}$, and thus we obtain a data set of n points in $\Re^{k-1}$, i.e., $\mathcal{V} = \{v_1, \ldots, v_n\}$. Then we perform

[5] To see this, let us first consider the case $d < n$. We can perform the SVD for the matrix $W_{\bar{S}}^T W_{\bar{S}} = V^T \mathrm{diag}\{\lambda_1, \ldots, \lambda_d\} V$, which takes $O(d^3)$ time. Then we can get the eigenvalues and their corresponding eigenvectors of $W_{\bar{S}}$ from the product $W_{\bar{S}} V$. The whole process takes only $O(d^3 + d^2 n)$ operations. The case for $d \geq n$ follows similarly.

the classical K-means clustering for the new reduced data set $\mathcal{V}$. From Theorem 2.1, this is equivalent to solving the following 0-1 SDP problem:

$$\begin{aligned} \min & \mathrm{Tr}\left((I-Z)\sum_{i=1}^{k-1}\lambda_i v^i (v^i)^T\right) \\ & Ze = e, \mathrm{Tr}(Z) = k, \\ & Z \geq 0, Z^2 = Z, Z = Z^T. \end{aligned} \tag{18}$$

For constrained K-means clustering, we need to solve the following subproblem:

$$\begin{aligned} \min & \mathrm{Tr}\left((I-Z)\sum_{i=1}^{k-1}\lambda_i v^i (v^i)^T\right) \\ & Z_{ii} \geq \frac{1}{\tilde{n}} \quad i = 1, \ldots, n, \\ & Ze = e, \mathrm{Tr}(Z) = k, \\ & Z \geq 0, Z^2 = Z, Z = Z^T. \end{aligned} \tag{19}$$

Finally, we partition all the entities in the original space based on the clustering on $\mathcal{V}$; i.e., the entities s_i, s_j belong to the same cluster if and only if v_i, v_j are in the same cluster.

The whole algorithm can be described as follows.

Approximation Algorithm 2.

Step 1. Calculate the projection of the matrix W onto the null space of e, i.e.,

$$W_1 = \left(I - \frac{1}{n}ee^T\right) W \left(I - \frac{1}{n}ee^T\right);$$

Step 2. Use the SVD method to compute the first $k-1$ largest eigenvalues of the matrix W_1 and their corresponding eigenvectors $v^1, \ldots, v^{k-1}$;

Step 3. Solve problem (18) (or (19)) for (constrained) K-means clustering;

Step 4. Assign all the entities in $\mathcal{S}$ based on the assignment obtained from Step 3.

Before closing this section, we discuss the complexity of Algorithm 2. We can resort to the exact algorithm in [7] or the algorithm in [18][6] to solve problem (18). According to Theorem 5 of [18], the algorithm takes $O(n^{(k-1)^2})$ time to find the global solution of the subproblem in Step 3 of Algorithm 2, which improves the running time when the same procedure is applied to solve the subproblem in [7]. This is because the working space in our algorithm is one dimension less than the space in [7]. In case of biclustering, the improvement is substantial since we can use our refined K-means described in the next section. For general kernel K-means clustering or spectral clustering, the overall complexity becomes $O(kn^2+n^{(k-1)^2})$. For the classical K-means clustering, the overall complexity is $O(\min\{n,d\}^3 + \min\{n,d\}mn + n^{(k-1)^2})$. In case of $k = 2$ and $d << n$, then the complexity of Algorithm 2 reduces to $O(n \log n)$. Since our algorithm uses only SVD and the constant in the big-O notation is very small, the algorithm for biclustering is very efficient and can be implemented easily. Because

[6]We point out that both works [7] and [18] employed the same technique for the MSSC model. However, in the present work, we cite only the results from [18] because the estimation of the complexity in [7] is not precise [6].

biclustering is the basis of the so-called divisive hierarchical clustering, our results are applicable to clustering methods based on divisive approaches.

We also point out that it is possible to use some approximation algorithm to solve the reduced problem in Algorithm 2. For example, if we apply the algorithm in [16] to the reduced problem, then we can find a 2-approximation to the reduced problem in $O(n^{k+1})$ time. From Theorem 4.1, we can immediately conclude that the obtained solution is a 4-approximation to the original problem. Note that the quality of the solution we have now is worse than what was obtained by applying the algorithm in [16] directly to the original data set. However, when $d \geq n$, the complexity of the algorithm in [16] goes up $O(n^{k+2})$. This implies that the complexity of the new approximation algorithm is better than the direct algorithm in [16] for scenarios like the kernel K-means clustering, spectral clustering, or problems from text mining where $d \geq n$.

4. Estimation of the approximate solution. In this section, we estimate the approximation solution provided by our algorithm. We first consider the case for the classical K-means clustering. It should be pointed out that in [7], Drineas et al. considered a similar algorithm based on the SVD of W and showed that their algorithm can provide a 2-approximation solution to the original K-means clustering. However, since the working subspace in our algorithm is different from that in [7], a new analysis is necessary to investigate the approximation ratio of the solution of Algorithm 2.

We first discuss the case of biclustering. One reason for this is that for biclustering, the subproblem involved in Step 3 of Algorithm 2 is in $\Re$. In such a case, the task in Step 3 of Algorithm 2 reduces to partitioning the data set $\mathcal{V} = \{v_i \in \Re, i = 1, \ldots, n\}$ into two clusters based on the MSSC model. Therefore, we can refer to the following refined K-means clustering in one dimension.

Refined K-means in One Dimension.

Step 0. Input the data set $\mathcal{V} = \{v_1, v_2, \ldots, v_n\}$;

Step 1. Sort the sequence so that

$$v_{i_1} \geq v_{i_2} \cdots \geq v_{i_n},$$

where $\{i_1, \ldots, i_n\}$ is a permutation of the index set $\{1, \ldots, n\}$;

Step 2. For $l = 1$ to n, set

$$C_1^l = \{v_{i_1}, \ldots, v_{i_l}\}, C_2^l = \{v_{i_{l+1}}, \ldots, v_{i_n}\},$$

and calculate the objective function

$$f(C_1^l, C_2^l) = \sum_{v_i \in C_1^l} \left(v_i - \frac{1}{l} \sum_{v_i \in C_1^l} v_i \right)^2 + \sum_{v_i \in C_2^l} \left(v_i - \frac{1}{n-l} \sum_{v_i \in C_2^l} v_i \right)^2$$

based on the partition (C_1^l, C_2^l);

Step 3. Find the optimal partition (C_1^*, C_2^*) such that

$$(C_1^*, C_2^*) = \arg \min_{l \in \{1, \ldots, n\}} f(C_1^l, C_2^l),$$

and output it as the final solution.

The above algorithm is similar to the algorithm in [15] for divisive k-clustering in low dimension. It is straightforward to see that for biclustering problems in $\Re$ based

on the MSSC model, the above procedure can find the global solution in $O(n \log n)$ time.

If $k \geq 3$, then we can use some existing exact algorithms in the literature for K-means clustering such as the algorithms in [7, 18] to solve the subproblem (18).

We next progress to estimate the approximation ratio of the solution of Algorithm 2. Let Z^* be a global solution to (8), and $\bar{Z}$ is the solution provided by Algorithm 2. Let us define

$$U = \frac{1}{n} ee^T + \sum_{i=1}^{k-1} v^i (v^i)^T. \tag{20}$$

It follows that

$$\mathrm{Tr}\left((I - U) \sum_{i=1}^{k-1} v^i (v^i)^T \right) = 0; \tag{21}$$

$$\mathrm{Tr}\left(U \sum_{i=k}^{n-1} v^i (v^i)^T \right) = 0. \tag{22}$$

From Theorem 3.1, we have

$$\mathrm{Tr}(W(I - Z^*)) \geq \mathrm{Tr}(W(I - U)). \tag{23}$$

It follows that

$$\mathrm{Tr}\big(W(I - \bar{Z})\big) = \mathrm{Tr}\big(W(I - U + U - \bar{Z})\big) \leq \mathrm{Tr}(W(I - Z^*)) + \mathrm{Tr}\big(W(U - \bar{Z})\big).$$

The above relation implies that if

$$\mathrm{Tr}\big(W(U - \bar{Z})\big) \leq \mathrm{Tr}(W(I - Z^*)), \tag{24}$$

then

$$\mathrm{Tr}\big(W(I - \bar{Z})\big) \leq 2\mathrm{Tr}(W(I - Z^*));$$

i.e., in the worst case, the solution provided by Algorithm 2 is a 2-approximation to the original K-means clustering.

In what follows we prove (24), which can be equivalently stated as

$$\mathrm{Tr}\big(W(I - Z^* + \bar{Z} - U)\big) \geq 0. \tag{25}$$

By the choices of $Z^*, \bar{Z}$, and U, it is easy to verify

$$(I - Z^* + \bar{Z} - U)e = 0, \tag{26}$$

$$\left(I - \frac{ee^T}{n}\right)(I - Z^* + \bar{Z} - U) = \left(I - \frac{ee^T}{n}\right)(I - Z^* + \bar{Z} - U)\left(I - \frac{ee^T}{n}\right). \tag{27}$$

It follows immediately that

$$\begin{aligned}
\mathrm{Tr}\big(W(I-Z^*+\bar{Z}-U)\big) &= \frac{1}{n}\mathrm{Tr}\big(W(I-Z^*+\bar{Z}-U)ee^T\big)+\mathrm{Tr}\big(W_1(I-Z^*+\bar{Z}-U)\big)\\
&= \mathrm{Tr}\left((I-Z^*+\bar{Z}-U)\sum_{i=1}^{n-1}\lambda_i v^i(v^i)^T\right)\\
&= \mathrm{Tr}\left((I-Z^*+\bar{Z}-U)\sum_{i=1}^{k-1}\lambda_i v^i(v^i)^T)\right)\\
&\quad +\mathrm{Tr}\left((I-Z^*+\bar{Z}-U)\sum_{i=k}^{n-1}\lambda_i v^i(v^i)^T\right)\\
&= \mathrm{Tr}\left((\bar{Z}-Z^*)\sum_{i=1}^{k-1}\lambda_i v^i(v^i)^T)\right)+Tr\left(I-Z^*+\bar{Z})\sum_{i=k}^{n-1}\right.\\
&\quad \left.\lambda_i v^i(v^i)^T\right)\\
&\geq \mathrm{Tr}\left((\bar{Z}-Z^*)\sum_{i=1}^{k-1}\lambda_i v^i(v^i)^T)\right),
\end{aligned}$$

where the last equality is given by (21) and (22), and the last inequality is implied by the fact that the matrix $I-Z^*+\bar{Z}$ is positive semidefinite. Recall that $\bar{Z}$ is the global solution of subproblem (18) and Z^* is only a feasible solution of (18); we therefore have

$$\mathrm{Tr}\left((\bar{Z}-Z^*)\sum_{i=1}^{k-1}\lambda_i v^i(v^i)^T)\right)\geq 0,$$

which further implies (24).

Now we are ready to state the main result in this section, which follows immediately from (23) and (24).

THEOREM 4.1. *Suppose that Z^* is a global solution to problem* (8) *and $\bar{Z}$ is the solution provided by Algorithm* 2*. Then, we have*

$$\mathrm{Tr}\big(W(I-\bar{Z})\big)\leq 2\mathrm{Tr}(W(I-Z^*)).$$

In what follows we estimate the approximation ratio of Algorithm 2 for constrained K-means clustering. It is worthwhile mentioning that in such a case, no polynomial algorithm has been reported in the literature to find a global solution of subproblem (19). However, suppose a global solution to (19) can be located; then by following a similar chain of reasoning as in the proof of Theorem 4.1, we can prove the following result.

THEOREM 4.2. *Suppose that Z^* is a global solution to problem* (12) *and $\bar{Z}$ is the solution provided by Algorithm* 2*. Then, we have*

$$\mathrm{Tr}\big(W(I-\bar{Z})\big)\leq 2\mathrm{Tr}(W(I-Z^*)).$$

5. Numerical experiments. To test the new algorithm, we have done some preliminary numerical experiments on several data sets from the UCI Machine Learning Repository[7] and internet newsgroups. All the experiments are done by using

[7] http://www.ics.uci.edu/~mlearn/MLRepository.html

TABLE 5.1
Results on three UCI data sets.

Data set	Stage 1	Stage 2	Global opt.
Soybean	404.4593	404.4593	404.4593
The Späth's	$6.0255e+11$	$6.0255e+11$	$6.0255e+11$
Spam e-mail	$9.43479784e+08$	$9.43479784e+08$	$9.43479784e+08$

MATLAB on a personal computer with a Pentium 4 1700 MHz Processor and a 256M memory. The power method is applied for calculating the largest eigenvalue and eigenvector for the matrix [11].

It should be mentioned that although the subproblem (18) can be solved by using the procedure in [7], the running time of the procedure is clearly too much for a reasonably large data set. Due to this fact, in our experiments, we restrict ourselves only to biclustering ($k = 2$).

We mention that in the following tables, we concentrate mainly on the quality of the obtained solutions. This is due to the fact that our tests are on biclustering problems. Based on our theoretical analysis, our algorithm enjoys the best complexity for the underlying problems compared with the algorithms in [16] and [36]. Even for the largest test problem (e-mail spam database) in this work, our algorithm takes less than one second to find the solution, while such a problem cannot be handled by using the LP relaxation in [36] because of the huge number (n^3) of constraints introduced in the LP model.[8]

Data sets from the UCI machine learning repository.

- *Soybean data set (small).* See also [32]. This data set has 47 points in $\Re^{35}$.
- *The Späth's postal zones.* This data set is from [41] about the post zones in Bavaria. It has 89 points in $\Re^3$.
- *Spam e-mail database.* Created by M. Hopkins et al. from Hewlett–Packard Labs. It has 4601 in $\Re^{57}$. For purposes of clustering, we have removed the last boolean attribute which indicates whether the e-mail was considered spam or not.

In our experiments, we use a two-phase strategy. After we obtain the partition of the data sets from Approximation Algorithm 2, we use the classical K-means to further improve the partition. In other words, we use Algorithm 2 only as a starting strategy for K-means clustering. In the following tables, we list the solutions from both phase 1 and phase 2.

Since, for the first two data sets, the global optimum has already been reported in [36] by using a linear programming model in the case of $K = 2$, we list it in the Global opt. column as a reference. The global solution for the third data set has been reported in [37]. The numerical results for general biclustering are summarized in Table 5.1. As one can see from the table, for the test problems, the solutions from stage 1 match the global solution.

Numerical results for balanced biclustering. We also test our algorithm for balanced biclustering. To find a global solution to balanced biclustering, we adapt the

[8] In one of our recent works [38], we also compared the algorithm in [16] with the so-called Q-means developed in [38] and our preliminary tests indicated that the Q-means algorithm outperformed the algorithm in [16]. For example, for the Email spam database under the same computational environment as in the present work, the Q-means took more than 300 seconds while the algorithm in [16] took more than half an hour to find a solution [38].

TABLE 5.2
Results for balanced biclustering.

Data set	Stage 1	Stage 2	LP/Q-means
Soybean	419.5473	418.5273	418.5273
The Späth's	$1.6310e+012$	$1.6310e+012$	$1.6310e+012$
Spam e-mail	$1.4049e+09$	$1.4046e+09$	$1.4046e+09$

TABLE 5.3
Newsgroups and their labels.

NG1	alt.atheism	**NG11**	rec.sport.hockey
NG2	comp.graphics	**NG12**	sci.crypt
NG3	comp.os.ms-windows.misc	**NG13**	sci.electronics
NG4	comp.sys.ibm.pc.hardware	**NG14**	sci.med
NG5	comp.sys.mac.hardware	**NG15**	sci.space
NG6	comp.windows.x	**NG16**	soc.religion.christian
NG7	misc.forsale	**NG17**	talk.politics.guns
NG8	rec.autos	**NG18**	talk.politics.mideast
NG9	rec.motorcycles	**NG19**	talk.politics.misc
NG10	rec.sport.baseball	**NG20**	talk.religion.misc

LP model in [36] slightly to incorporate balanced constraints. The solution obtained from the LP model gives us a lower bound for the global optimum of the balanced biclustering. We also pointed out that for the third data set, its relatively large size prevents us from the use of the LP model due to the enormous amount $O(n^3)$ of constraints involved in the model. In such a case, we list the result from [38], which is derived by a so-called Q-means heuristic for the same data and same balanced constraint.

In the experiments for the last two data sets, we require that each cluster have at least $n/3$ entities. For the soybean data set, we require that each cluster have at least 22 entities. This is because the data set itself is fairly balanced already (the optimal biclustering has a $(20, 27)$ distribution). Table 5.2 summarizes the results.

From the above tables we can see that the solution from phase 1 is very close to the solution from phase 2. In all the cases, the solution from phase 2 matches the global solution of the underlying problem.

Internet newsgroups. Text mining has been popular in document analysis, search engine, and knowledge discovery in a large volume of text data. We have also performed experiments on newsgroup articles submitted to 20 newsgroups.[9] This data set has also been used in [5, 13, 45], where a similar framework to ours was used to solve the problem. The algorithm we use is still the two-phase heuristic which was introduced in the last section.

This data set consists of about 20,000 articles (e-mail messages) evenly distributed among the 20 newsgroups. We list the name of the newsgroups together with the associated group labels.

Before constructing the word-document matrices, we perform the preprocessing by using the *bow* toolkit, a preprocessor similar to what was employed in [5, 13, 45]. In particular, we use the tokenization option such that the UseNet headers are stripped, since the headers include the name of the correct newsgroup, and we also

[9]The news group data together with the bow toolkit for preprocessing can be downloaded from http://www-2.cs.cmu.edu/afs/cs/project/theo-11/www/naive-bayes.html.

TABLE 5.4
Results on internet newsgroup data sets.

Data set	Stage 1	Stage 2	LP
NG1/NG2	92.6690	92.6630	92.6630
NG2/NG3	94.0377	94.0377	94.0377
NG8/NG9	93.7051	93.5380	93.4007
NG10/NG11	92.0785	92.0299	92.0299
NG1/NG15	91.9277	91.9011	91.9011
NG18/NG19	92.2275	92.2035	92.2035

apply stemming [27]. Afterward, we apply the standard tf.idf weighting scheme and normalize each document vector to have unit Euclidean length. Finally, we conduct feature selection where 500 words are selected according to the mutual information between words and documents in an unsupervised manner.

In our experiment, we choose 50 random document vectors from each of two newsgroups. This implies that for each test, we have a data set with 100 points in $\Re^{500}$. Then we apply our approximation algorithm to the problem. The results are summarized in Table 5.3. Note that, since the global optima are not known for these data sets, we use the linear programming relaxation model proposed in [36] to get a lower bound on the global optimum. More specifically, we implement the LP relaxation model (14) in [36] using package CPLEX 7.1 with AMPL interface on an IBM RS-6000; by solving this LP problem, we can obtain a lower bound for the global optimum solution. Apparently, if the solution obtained from the LP relaxation equals to the solution provided by our two-phase heuristic, then it must be a global optimal solution of the original problem. The numerical results for the internet newsgroups are summarized in Table 5.4.

From the above experiments, we can conclude that our deterministic two-phase heuristic performs very well on these data sets and it finds the global optimum for most of these data sets.

6. Conclusions. In this paper, we reformulated the classical MSSC as a 0-1 SDP. Our new model not only provides a unified framework for several existing clustering approaches, but also opens new avenues for clustering. An approximation method based on the SDP relaxation and PCA has been proposed to attack the underlying 0-1 SDP. It is shown that in the worst case, our method can provide a 2-approximate solution to the original classical or constrained K-means clustering. Preliminary numerical tests indicate that our algorithm can always find a global solution for biclustering.

Several important issues regarding the new framework remain open. First, for general $k \geq 3$, although subproblem (18) can be solved by using some exact algorithms in the literature [7, 18], its complexity is still exponential in k. This makes the algorithm impractical for relatively large data sets. Second, the current model can deal with only a simple case of constrained K-means clustering. The issue of how to deal with general constrained K-means clustering still remains open. More study is needed to address these questions.

Acknowledgments. The authors would like to thank Dr. Y. Xia, Prof. F. Rendl, and Prof. K. C. Toh for their useful advice in the preparation of this paper.

REFERENCES

[1] P. K. Agarwal and Procopiuc, *Exact and approximation algorithms for clustering*, Algorithmica, 33 (2002), pp. 201–226.

[2] F. R. Bach and M. I. Jordan, *Learning Spectral Clustering*, Adv. Neural Inf. Process. Syst., 16, MIT Press, Cambridge, MA, 2004.

[3] P. Bradley, K. Bennet, and A. Demiriz, *Constrained K-means Clustering*, Tech. report MSR-TR-2000-65, Microsoft Research, 2000.

[4] P. S. Bradley, U. M. Fayyad, and O. L. Mangasarian, *Mathematical programming for data mining: Formulations and challenges*, INFORMS J. Comput., 11 (1999), pp. 217–238.

[5] C. Ding and X. He, *K-means clustering via principal component analysis*, in Proceedings of the 21st Annual International Conference on Machine Learning, Banff, Canada, 2004.

[6] P. Drineas, *Private communication*, 2005.

[7] P. Drineas, A. Frieze, R. Kannan, R. Vempala, and V. Vinay, *Clustering large graphs via singular value decomposition*, Machine Learning, 56 (2004), pp. 9–33.

[8] O. Du Merle, P. Hansen, B. Jaumard, and N. Mladenović, *An interior-point algorithm for minimum sum-of-squares clustering*, SIAM J. Sci. Comput., 21 (2000), pp. 1485–1505.

[9] J. Ghosh, *Scalable Clustering*, in The Handbook of Data Mining, N. Ye, ed., Lawrence Erlbaum Associate, Inc., 2003, pp. 247–277.

[10] M. X. Goemans, *Semidefinite programming in combinatorial optimization*, Math. Programming, 79 (1997), pp. 143–161.

[11] G. Golub and C. V. Loan, *Matrix Computations*, John Hopkins University Press, Baltimore, MD, 1996.

[12] A. D. Gordon and J. T. Henderson, *An algorithm for Euclidean sum of squares classification*, Biometrics, 33 (1977), pp. 355–362.

[13] M. Gu, H. Zha, C. Ding, X. He, and H. Simon, *Spectral Relaxation Models and Structure Analysis for k-way Graph Clustering and Bi-clustering*, Penn State University Technical Report, State College, PA, 2001.

[14] P. Hansen and B. Jaumard, *Cluster analysis and mathematical programming*, Math. Programming, 79(B) (1997), pp. 191–215.

[15] P. Hansen, B. Jaumard, and N. Mladenović, *Minimum sum of squares clustering in a low dimensional space*, J. Classification, 15 (1998), pp. 37–55.

[16] S. Hasegawa, H. Imai, M. Inaba, N. Katoh, and J. Nakano, *Efficient algorithms for variance-based k-clustering*, in Proceedings of the First Annual Pacific Conference on Computer Graphics and Applications (Seoul, Korea), Vol. 1, World Scientific, Singapore, 1993, pp. 75–89.

[17] H. Howard, *Classifying a population into homogeneous groups*, in Operational Research in Social Science, J. R. Lawrence, ed., Tavistock Publications, London, 1966.

[18] M. Inaba, N. Katoh, and H. Imai, *Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering: Extended abstract*, in Proceedings of the Tenth Annual Symposium on Computational Geometry, ACM, New York, 1994, pp. 332–339.

[19] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, Prentice–Hall, Englewood Cliffs, NJ, 1988.

[20] A. K. Jain, M. N. Murty, and P. J. Flynn, *Data clustering: A review*, ACM Computing Surveys, 31 (1999), pp. 264–323.

[21] R. C. Jancey, *Multidimensional group analysis*, Australian J. Botany, 14 (1966), pp. 127–130.

[22] I. Jolliffe, *Principal Component Analysis*, 2nd ed., Springer-Verlag, New York, 2002.

[23] L. Kaufman and P. P. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley, New York, 1990.

[24] S. E. Karisch and F. Rendl, *Semidefinite programming and graph equipartition*, Fields Inst. Commun., 18 (1998), pp. 77–95.

[25] A. Kumar, Y. Sabharwal, and S. Sen, *A simple linear time* $(1+\epsilon)$*-approximation algorithm for k-means clustering in any dimensions*, in Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science (FOCS'04), IEEE Computer Society, Los Alamitos, CA, 2004, pp. 454–462.

[26] A. Leisser and F. Rendl, *Graph partitioning using linear and semidefinite programming*, Math. Program. (B), 95 (2003), pp. 91–101.

[27] A. McCallum, *Bow: A Toolkit for Statistical Language Modeling, Text Retrieval, Classification, and Clustering*, available online at http://www.cs.umass.edu/~mccallum/bow/.

[28] J. McQueen, *Some methods for classification and analysis of multivariate observations*, Computer and Chemistry, 4 (1967), pp. 257–272.

[29] O. L. Mangasarian, *Nonlinear Programming,* Classics Appl. Math. 10, SIAM, Philadelphia, 1994.
[30] O. L. Mangasarian, *Mathematical programming in data mining*, Data Min. Knowl. Discov., 1 (1997), pp. 183–201.
[31] M. Meila and J. Shi, *A random walks view of spectral segmentation*, International Workshop on AI and Stat., (2001).
[32] R. S. Michalski and R. L. Chilausky, *Learning by being told and learning from examples: An experimental comparison of the two methods of knowledge acquisition in the context of developing an expert system for soybean disease diagnosis*, International Journal of Policy Analysis and Information Systems, 4 (1980), pp. 125–161.
[33] J. Matousek, *On approximate geometric k-clustering*, Discrete Comput. Geom., 24 (2000), pp. 61–84.
[34] A. Y. Ng, M. I. Jordan, and Y. Weiss, *On spectral clustering: Analysis and an algorithm*, Proc. Neural Info. Processing Systems, NIPS, 14 (2001).
[35] M. L. Overton and R. S. Womersley, *Optimality conditions and duality theory for minimizing sums of the largest eigenvalues of symmetric matrices*, Math. Programming, 62 (1993), pp. 321–357.
[36] J. M. Peng and Y. Xia, *A new theoretical framework for K-means clustering*, in Foundation and Recent Advances in Data Mining, W. Chu and T. Lin, eds., Springer-Verlag, New York, 2005, pp. 79–98.
[37] J. M. Peng and Y. Xia, *A cutting plane method for the minimum-sum-of-squared error clustering*, in Proceedings of the SIAM International Conference on Data Mining, Newport Beach, CA, 2005.
[38] G. Ma, J. Peng, and Y. Wei, *On approximate balanced bi-clustering*, in Computing and Combinatorics Conference, Lecture Notes in Comput. Sci. 3595, Springer-Verlag, Berlin, 2005, pp. 661–670.
[39] E. H. Ruspini, *Numerical methods for fuzzy clustering*, Inform. Sci., 2 (1970), pp. 319–350.
[40] J. Shi and J. Malik, *Normalized cuts and image segmentation*, IEEE. Trans. Pattern Analysis and Machine Intelligence, 22 (2000), pp. 888–905.
[41] H. Späth, *Cluster Analysis Algorithms for Data Reduction and Classification of Objects*, Ellis Horwood Ltd., West Sussex, UK, 1980.
[42] J. H. Ward, *Hierarchical grouping to optimize an objective function*, J. Amer. Statist. Assoc., 58 (1963), pp. 236–244.
[43] Y. Weiss, *Segmentation using eigenvectors: A unifying view*, in Proceedings of the IEEE International Conference on Computer Vision, IEEE Computer Society, Los Alamitos, CA, 1999, pp. 975–982.
[44] E. P. Xing and M. I. Jordan, *On Semidefinite Relaxation for Normalized k-cut and Connections to Spectral Clustering*, Tech. report CSD-03-1265, UC Berkeley, Berkeley, CA, 2003.
[45] H. Zha, C. Ding, M. Gu, X. He, and H. Simon, *Spectral relaxation for K-means clustering*, in Advances in Neural Information Processing Systems 14, T. Dietterich, S. Becker, and Z. Ghahramani, eds., MIT Press, Cambridge, MA, 2002, pp. 1057–1064.

SIAM J. OPTIM.
Vol. 18, No. 1, pp. 206–222

DECOMPOSITION-BASED INTERIOR POINT METHODS FOR TWO-STAGE STOCHASTIC SEMIDEFINITE PROGRAMMING*

SANJAY MEHROTRA† AND M. GÖKHAN ÖZEVIN‡

Abstract. We introduce two-stage stochastic semidefinite programs with recourse and present an interior point algorithm for solving these problems using Bender's decomposition. This decomposition algorithm and its analysis extend Zhao's results [*Math. Program.*, 90 (2001), pp. 507–536] for stochastic linear programs. The convergence results are proved by showing that the logarithmic barrier associated with the recourse function of two-stage stochastic semidefinite programs with recourse is a strongly self-concordant barrier on the first stage solutions. The short-step variant of the algorithm requires $O(\sqrt{p+Kr}\ln\mu^0/\epsilon)$ Newton iterations to follow the first stage central path from a starting value of the barrier parameter μ^0 to a terminating value ϵ. The long-step variant requires $O((p+Kr)\ln\mu^0/\epsilon)$ damped Newton iterations. The calculation of the gradient and Hessian of the recourse function and the first stage Newton direction decomposes across the second stage scenarios.

Key words. stochastic programming, semidefinite programming, Benders decomposition, interior point methods, primal methods

AMS subject classifications. 90C05, 90C15, 90C20, 90C22, 90C25

DOI. 10.1137/050622067

1. Introduction.

We introduce and study the two-stage stochastic semidefinite programming (TSSDP) problem with recourse in the dual standard form:

$$\begin{aligned} \max \quad & \eta(x) := c^T x + \varrho(x) \\ \text{s.t.} \quad & Ax + s = b, \\ & s \in \mathcal{K}^p, \end{aligned} \tag{1.1}$$

where

$$\varrho(x) := E[\varrho^{\tilde{\xi}}(x)] \tag{1.2}$$

and

$$\begin{aligned} \varrho^{\xi}(x) := \max \quad & {d^{\xi}}^T y^{\xi} \\ \text{s.t.} \quad & W^{\xi} y^{\xi} + s^{\xi} = h^{\xi} - T^{\xi} x, \\ & s^{\xi} \in \mathcal{K}^r. \end{aligned} \tag{1.3}$$

In the first stage problem (1.1), $x \in \mathbb{R}^n$ and $s \in \mathbb{R}^{p^2}$ are decision variables. A is a $p^2 \times n$ matrix with n linearly independent columns that are obtained by vectorization of n symmetric real $p \times p$ matrices and $b \in \mathbb{R}^{p^2}$. We have chosen this form of TSSDP

*Received by the editors January 5, 2005; accepted for publication (in revised form) September 12, 2006; published electronically March 19, 2007. The research of both authors was supported in part by NSF-DMI-0200151 and ONR-N00014-01-1-0048. An earlier draft of this paper appeared under the title "Two-Stage Stochastic Semidefinite Programming and Decomposition Based Interior Point Methods: Theory," IEMS Technical Report 2004-16, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL, 2004. The work was performed while the second author was at Northwestern University.

http://www.siam.org/journals/siopt/18-1/62206.html

†Corresponding author. Department of Industrial Engineering & Management Sciences, Northwestern University, Evanston, IL 60208 (mehrotra@northwestern.edu).

‡ZS Associates, 1800 Sherman Avenue, Evanston, IL 60201 (gokhan.ozevin@zsassociates.com).

for notational convenience in the analysis of this paper. By $S := \mathbf{mat}(s)$ we denote the $\nu \times \nu$ matrix whose (i,j)th element is the $((j-1)\nu+i)$th element of a vector $s \in \mathbb{R}^{\nu^2}$. We use $\mathbf{vec}(S)$ to denote a vector whose $((j-1)\nu+i)$th element is the (i,j)th element of a matrix S. The cone $\mathcal{K}^\nu := \{\mathbf{vec}(S) \mid S \in \mathbb{R}^{\nu\times\nu} \text{is symmetric positive semidefinite}\}$ is the cone of vectors obtained from the vectorization of symmetric positive semidefinite matrices. $\mathcal{K}^\nu_+$ is used to describe the cone generated by positive definite matrices.

Randomness in the second stage is governed by the random variable $\tilde{\xi}$. We assume that the support Ξ of $\tilde{\xi}$ is discrete and finite. $E[\cdot]$ in (1.2) represents the expectation. For each realization ξ of $\tilde{\xi}$, $y^\xi \in \mathbb{R}^m$ and $s^\xi \in \mathbb{R}^{r^2}$ are decision variables. $h^\xi \in \mathbb{R}^{r^2}$ and T^ξ is a $r^2 \times n$ matrix with n linearly independent columns that are obtained by vectorization of n symmetric real $r \times r$ matrices. Similarly, W^ξ is a $r^2 \times m$ matrix with m linearly independent columns that are obtained by vectorization of m symmetric real $r \times r$ matrices. Under suitable assumptions (see section 2), $\varrho^\xi(x)$ is finite and well defined for all feasible first stage solutions (x,s). $\varrho(x)$ is called the recourse function.

The TSSDP problem is a natural generalization of a semidefinite programming problem [14] to its two-stage stochastic programming counterpart. Problems where objective and constraints are defined by convex quadratic inequalities or second order cone inequalities are special cases. The linear-quadratic model introduced by Rockafellar and Wets [11] is also a special case. In general, we can write the explicit extensive formulation of these problems as a large-scale semidefinite program. We can then solve this extensive formulation directly, particularly by using primal-dual interior point methods, exploiting its special structure through efficient matrix factorization schemes [3, 4, 5]. However, the focus of this paper is in developing decomposition-based interior point methods for TSSDP in the spirit of Bender's decomposition.

Optimal dual solutions of second stage problems are used in the evaluation of the gradient and Hessian of the recourse function, which is central to the algorithm. In practice, we can calculate the gradient and Hessian information approximately. This view of the algorithm has several potential advantages since it does not require explicit knowledge of all the scenarios and associated variables in the algorithm up front. First, the scenarios can be added as the algorithm progresses. This has the potential for speeding up the algorithm in its early stages. Mehrotra and Özevin [7] provide evidence for this benefit. Second, if computations for some of the scenarios fail due to unreliability of available computational resources, one may still be able to proceed with the algorithm. This allows for implementations in a distributed computing environment where some of the computing nodes may not be reliable.

In general, the recourse function $\varrho(x)$ is not differentiable with respect to x everywhere. The decomposition approaches either use the nonsmooth optimization techniques [1, 2, 13], or use techniques to smooth this function [11, 12]. Given the success of interior point methods, it is logical to investigate whether decomposition-based interior point algorithms are possible for stochastic programming problems. Zhao [15] developed an interior decomposition algorithm for linear two-stage stochastic programs by regularizing the second stage problem with a log barrier. In particular, he showed that the log barrier associated with the recourse function of two-stage stochastic linear programs behaves as a strongly self-concordant barrier (see Nesterov and Nemirovskii [8] and Renegar [10]) on the first stage solutions. Mehrotra and Ozevin [6] extended Zhao's analysis for two-stage stochastic convex quadratic programs. In this paper we show that the recourse function is also strongly self-concordant for TSSDP. This allows us to give a Benders decomposition-based linearly convergent interior point algorithm for TSSDP. The convergence analysis of this paper provides

the conceptual framework for a more practical algorithm developed and implemented in [7].

This paper is organized as follows. In section 2 we give barrier problem formulations and our assumptions. In section 3 we show that the barrier recourse function comprises a self-concordant family. In section 4 we present short- and long-step variants of an interior point decomposition algorithm and state the convergence results. Proofs of these convergence theorems are given in section 5.

We use the following additional notation. For any strictly positive vector x in $\mathbb{R}^n$, we define $x^{-1} := (x_1^{-1}, \ldots, x_n^{-1})^T$. An identity matrix of appropriate dimension is denoted by I. Throughout this paper we use "∇","∇^2","∇^3" to denote the gradient, Hessian, and the third order derivative with respect to x, and a " $'$ " for the derivative with respect to a single variables other than x. For example,

$$[\{\nabla^2 f(\mu, x)\}']_{i,j} = \frac{\partial}{\partial \mu}\left(\frac{\partial^2 f(\mu, x)}{\partial x_i \partial x_j}\right).$$

"∇" is also used to denote the Jacobian of a vector function. $A \otimes B$ represents the Kronecker product of matrices A and B. The Kronecker product satisfies relationship $[A \otimes B][C \otimes D] = [AC \otimes BD]$, assuming that the number of rows in A and B equals the number of columns in C and D. Also, $(A \otimes B)\mathbf{vec}(C) = \mathbf{vec}(BCA^T)$.

2. Problem formulation and assumptions. Let the random variable $\tilde{\xi}$ have a finite discrete support $\Xi = \{\xi^1, \ldots, \xi^K\}$ with probabilities $\{\pi^1, \ldots, \pi^K\}$. For simplicity of notation we define $\varrho^i(x) := \varrho^{\xi^i}(x)$, $T^i := T^{\xi^i}$, $W^i := W^{\xi^i}$, $h^i := h^{\xi^i}$, $y^i := y^{\xi^i}$, and $d^i := \pi^i d^{\xi^i}$. The problem (1.1)–(1.3) is rewritten as

$$\begin{aligned} \max \quad & \eta(x) := c^T x + \varrho(x) \\ \text{s.t.} \quad & Ax + s = b, \\ & s \in \mathcal{K}^p, \end{aligned} \tag{2.1}$$

where

$$\varrho(x) := \sum_{i=1}^{K} \varrho^i(x), \tag{2.2}$$

and for $i = 1, \ldots, K$,

$$\begin{aligned} \varrho^i(x) := \max \quad & {d^i}^T y^i \\ \text{s.t.} \quad & W^i y^i + s^i = h^i - T^i x, \\ & s^i \in \mathcal{K}^r. \end{aligned} \tag{2.3}$$

Let γ and λ^i be the first and second stage dual multipliers. The dual of (2.3) is

$$\begin{aligned} \min \quad & (h^i - T^i x)^T \lambda^i \\ \text{s.t.} \quad & {W^i}^T \lambda^i = d^i, \\ & \lambda^i \in \mathcal{K}^r. \end{aligned} \tag{2.4}$$

Here $s^i \in \mathbb{R}^{r^2}$, $W^i \in \mathbb{R}^{r^2 \times m}$, and h^i, T^i is data of appropriate dimensions.

Let us define the following feasibility sets:

$$\mathcal{F}^i(x) := \{y^i \mid W^i y^i + s^i = h^i - T^i x,\ s^i \in \mathcal{K}^r\}, \qquad \mathcal{F}_1^i := \{x \mid \mathcal{F}^i(x) \neq \emptyset\},$$
$$\mathcal{F}_1 := \cap_{i=1}^K \mathcal{F}_1^i, \qquad \mathcal{F}_0 := \mathcal{F}_1 \cap \{x \mid Ax + s = b,\ s \in \mathcal{K}^p\}, \text{ and}$$
$$\mathcal{F} := \{(x, s, \gamma) \times (y^1, s^1, \lambda^1, \ldots, y^K, s^K, \lambda^K) \mid Ax + s = b, s \in \mathcal{K}^p; W^i y^i + s^i = h^i - T^i x,$$
$$s^i \in \mathcal{K}^r; {W^i}^T \lambda^i = d^i, \lambda^i \in \mathcal{K}^r, \text{ for } i = 1, \ldots, K;\ A^T\gamma + \textstyle\sum_{i=1}^K {T^i}^T \lambda^i = c\}.$$

We make the following assumptions:

A1. The set $\mathcal{F}$ is not empty, is bounded, and has a nonempty relative interior.

A2. Matrices A and W^i have full column rank.

Assumption A1 requires that primal and dual feasible sets of the explicit deterministic equivalent formulation of (2.1)–(2.3) have nonempty interiors. In particular, it assumes strong duality (see, for example, Ramana, Tunçel, and Wolkowicz [9]) for first and second stage semidefinite programs. This assumption also ensures that the recourse function $\varrho(x) : \mathcal{F} \to \mathbb{R}$ is finite and well defined. In practice this assumption can be ensured by introducing artificial variables. Assumption A2 is for convenience.

Consider the following log-barrier decomposition problem

$$\begin{aligned} \max \quad & \eta(\mu, x) := c^T x + \rho(\mu, x) + \mu \ln\det S \\ \text{s.t.} \quad & Ax + s = b, \\ & s \in \mathcal{K}^p, \end{aligned} \tag{2.5}$$

where

$$\rho(\mu, x) := \sum_{i=1}^K \rho^i(\mu, x) \tag{2.6}$$

and for $i = 1, \ldots, K$

$$\begin{aligned} \rho^i(\mu, x) := \max \quad & {d^i}^T y^i + \mu \ln\det S^i \\ \text{s.t.} \quad & W^i y^i + s^i = h^i - T^i x, \\ & s^i \in \mathcal{K}^r. \end{aligned} \tag{2.7}$$

The log-barrier problem associated with the dual (2.4) is given by

$$\begin{aligned} \min \quad & (h^i - T^i x)^T \lambda^i - \mu \ln\det \Lambda^i \\ \text{s.t.} \quad & {W^i}^T \lambda^i = d^i, \\ & \lambda^i \in \mathcal{K}^r. \end{aligned} \tag{2.8}$$

Note that, for a given $\mu > 0$, the log-barrier recourse function $\rho(\mu, x) < \infty$ iff $x \in \mathcal{F}_1$. Hence, it describes the interior of $\mathcal{F}_0$ implicitly. Assumption A1 implies that problems (2.5) and (2.7)–(2.10) have a unique solution. Since the objective in problems (2.7) and (2.8), is respectively, concave and convex function, (y^i, s^i) and λ^i are optimal solutions to (2.7) and (2.8), respectively, iff they satisfy the following optimality conditions:

$$\begin{aligned} & {W^i}^T \lambda^i = d^i, \\ & W^i y^i + s^i = h^i - T^i x, \\ & S^i \Lambda^i = \mu I, \\ & \lambda^i, s^i \in \mathcal{K}_+^r, \end{aligned} \tag{2.9}$$

where $\Lambda^i = \mathbf{mat}(\lambda^i)$. Throughout this paper we denote the optimal solution of the first stage problem (2.5) by $x(\mu)$, and the solutions of the optimality conditions (2.9) for a given $x \in \mathcal{F}_1$ by $(y^i(\mu,x), s^i(\mu,x), \lambda^i(\mu,x))$. The optimal solutions of (2.5)–(2.7) and those of the log-barrier problem

$$\begin{aligned} \max \quad & c^T x + \sum_{i=1}^{K} {d^i}^T y^i + \mu \ln\det S + \mu \sum_{i=1}^{K} \ln\det S^i \\ \text{s.t.} \quad & Ax + s = b, \\ & W^i y^i + s^i = h^i - T^i x, \quad i = 1,\dots,K, \\ & s \in \mathcal{K}^p,\ s^i \in \mathcal{K}^r, \quad i = 1,\dots,K, \end{aligned} \tag{2.10}$$

associated with the extensive formulation of (2.1)–(2.3) have the following relationship.

PROPOSITION 2.1. *For a given $\mu > 0$, if $(x(\mu), s(\mu); y^1(\mu), s^1(\mu), \dots, y^K(\mu), s^K(\mu))$ is the optimal solution of* (2.10)*, then $(x(\mu), s(\mu))$ is the optimal solution of* (2.5)*, and $(y^1(\mu), s^1(\mu), \dots, y^K(\mu), s^K(\mu))$ are the optimal solutions of subproblems* (2.7) *for the given μ and $x = x(\mu)$. Conversely, if for a given μ, $(x(\mu), s(\mu))$ is the optimal solution of* (2.5) *and $(y^1(\mu), s^1(\mu), \dots, y^K(\mu), s^K(\mu))$ are the optimal solutions of* (2.7) *with $x = x(\mu)$, then $(x(\mu), s(\mu) y^1(\mu), s^1(\mu), \dots, y^K(\mu), s^K(\mu))$ is the optimal solution of* (2.10).

3. The self-concordance properties of the log-barrier recourse.

3.1. Computation of $\nabla\eta(\mu,x)$ and $\nabla^2\eta(\mu,x)$. From (2.9) we can show that the optimal objective values of primal and dual barrier problems (2.7)–(2.8) differ by a constant term, in particular

$$\rho^i(\mu,x) = (h^i - T^i x)\lambda^i(\mu,x) - \mu \ln\det \Lambda^i(\mu,x) + r\mu(1 - \ln\mu). \tag{3.1}$$

In order to compute $\nabla\eta(\mu,x)$ and $\nabla^2\eta(\mu,x)$ we need to determine the derivative of $\lambda^i(\mu,x)$ with respect to x. Let $(y^i, \lambda^i, s^i) := (y^i(\mu,x), \lambda^i(\mu,x), s^i(\mu,x))$. Differentiating (2.9) with respect to x, we obtain

$$\begin{aligned} {W^i}^T \nabla\lambda^i &= 0, \\ W^i \nabla y^i + \nabla s^i &= -T^i, \\ (I \otimes S^i)\nabla\lambda^i + (\Lambda^i \otimes I)\nabla s^i &= 0. \end{aligned} \tag{3.2}$$

Solving the system (3.2), we get

$$\begin{aligned} \nabla y^i &= -{R^i}^{-1} {W^i}^T {Q^i}^2 T^i, \\ \nabla\lambda^i &= Q^i P^i Q^i T^i, \\ \nabla s^i &= -{Q^i}^{-1} P^i Q^i T^i, \end{aligned} \tag{3.3}$$

where

$$Q^i := Q^i(\mu,x) = (\Lambda^i \otimes {S^i}^{-1})^{1/2}, \qquad R^i := R^i(\mu,x) = {W^i}^T {Q^i}^2 W^i \tag{3.4}$$

$$\text{and} \quad P^i := P^i(\mu,x) = I - Q^i W^i {R^i}^{-1} {W^i}^T Q^i. \tag{3.5}$$

Now differentiating (3.1) and using the optimality conditions (2.9) and (3.3), we can verify that

$$\nabla\rho^i(\mu,x) = -{T^i}^T\lambda^i(\mu,x) \quad \text{and} \quad \nabla^2\rho^i(\mu,x) = -{T^i}^T\nabla\lambda^i(\mu,x). \tag{3.6}$$

Hence,

$$\nabla\eta(\mu,x) = c - \sum_{i=1}^{K} {T^i}^T\lambda^i(\mu,x) - \mu A^T s^{-1}, \tag{3.7}$$

$$\nabla^2\eta(\mu,x) = -\sum_{i=1}^{K} {T^i}^T\nabla\lambda^i(\mu,x) - \mu A^T(S^{-1}\otimes S^{-1})A. \tag{3.8}$$

Then, substituting for $\nabla\lambda_i$ in (3.8), we get

$$\nabla^2\eta(\mu,x) = -\sum_{i=1}^{K} {T^i}^T Q^i P^i Q^i T^i - \mu A^T(S^{-1}\otimes S^{-1})A. \tag{3.9}$$

3.2. Self-concordance of the recourse function. The following definition of self-concordant functions was introduced by Nesterov and Nemirovskii [8].

DEFINITION 3.1 (Nesterov and Nemirovskii [8]). *Let $\mathcal{E}$ be a finite-dimensional real vector space, $\mathcal{Q}$ be an open nonempty convex subset of $\mathcal{E}$, and $f : \mathcal{Q} \to \mathbb{R}$ be a function, $\alpha > 0$. f is called α-self-concordant on $\mathcal{Q}$ with the parameter value α if $f \in C^3$ is a convex function on $\mathcal{Q}$, and, for all $x \in \mathcal{Q}$ and $h \in \mathcal{E}$, the following inequality holds:*

$$|\nabla^3 f(x)[h,h,h]| \le 2\alpha^{-1/2}(\nabla^2 f(x)[h,h])^{3/2}.$$

An α-self-concordant on $\mathcal{Q}$ function f is called strongly α-self-concordant on $\mathcal{Q}$ if $f(x^i)$ tends to infinity along every sequence $\{x^i \in \mathcal{Q}\}$ converging to a boundary point of $\mathcal{Q}$.

We now show that recourse function $\rho(\mu,x)$ behaves as a strongly self-concordant barrier on $\mathcal{F}_1$.

LEMMA 3.1. *For any fixed $\mu > 0$, $\rho^i(\mu,\cdot)$ is strongly μ-self-concordant on $\mathcal{F}_1^i, i = 1,\dots,K$.*

Proof. For any $\mu > 0$, $d \in \mathbb{R}^n$, and $\bar{x} \in \{x \mid \rho^i(x) < \infty\}$ we define the univariate function

$$\Phi^i(t) := \nabla^2\rho^i(\mu,\bar{x}+td)[d,d].$$

Note that $\Phi^i(0)' = \nabla^3\rho^i(\mu,\bar{x})[d,d,d]$. Along every sequence $\{x^j \in \mathcal{F}_1^i\}$ converging to the boundary of $\mathcal{F}_1^i$, $\rho^i(\mu,x^j)$ tends to infinity. To prove this lemma it suffices to show that

$$|\Phi^i(0)'| \le \frac{2}{\sqrt{\mu}}|\Phi^i(0)|^{3/2}.$$

Let $(\lambda^i(t), P^i(t), s^i(t), Q^i(t), R^i(t)) := (\lambda^i(\mu,\bar{x}+td), P^i(\mu,\bar{x}+td), s^i(\mu,\bar{x}+td), Q^i(\mu,\bar{x}+td), R^i(\mu,\bar{x}+td))$. We define $u^i(t) := P^i(t)Q^i(t)T^i(t)d$. The argument "$(t)$" is dropped when considering all of these variables and their derivatives at $t = 0$, e.g., $u' := u'(0)$. Note that $\Phi^i(0) = -{u^i}^T u^i = -\|u^i\|^2$ and thus $|\Phi^i(0)'| = |2{u^i}^T {u^i}'|$.

The first equality below follows from using (3.4)–(3.5). The second equality is derived by using derivatives by parts. The third equality uses ${R^i}' = {W^i}^T[Q^i{Q^i}' + {Q^i}'Q^i]W^i$.

We have

$$\begin{aligned}
u^{i\prime} &= [Q^i - Q^i W^i R^{i^{-1}} W^i Q^{i^2}]' T^i d \\
&= [Q^{i\prime} - Q^{i\prime} W^i R^{i^{-1}} W^i Q^{i^2} + Q^i W^i R^{i^{-1}} R^{i\prime} R^{i^{-1}} W^i Q^{i^2} \\
&\qquad - Q^i W^i R^{i^{-1}} W^i (Q^i Q^{i\prime} + Q^{i\prime} Q^i)] T^i d \\
&= [Q^{i\prime}(I - W^i R^{i^{-1}} W^i Q^{i^2}) - Q^i W^i R^{i^{-1}} W^i (Q^i Q^{i\prime} + Q^{i\prime} Q^i) \\
&\qquad (I - W^i R^{i^{-1}} W^i Q^{i^2})] T^i d \\
&= [(Q^{i\prime} - Q^i W^i R^{i^{-1}} W^i (Q^i Q^{i\prime} + Q^{i\prime} Q^i))](I - W^i R^{i^{-1}} W^i Q^{i^2}) T^i d \\
&= [(Q^{i\prime} - Q^i W^i R^{i^{-1}} W^i (Q^i Q^{i\prime} + Q^{i\prime} Q^i))] Q^{i^{-1}} u^i \\
&\qquad \text{(noting that } (I - W^i R^{i^{-1}} W^i Q^{i^2}) T^i d = Q^{i^{-1}} u^i). \qquad (3.10)
\end{aligned}$$

Observing that ${u^i}^T Q^i W^i = 0$, from (3.10) we get

$$\begin{aligned}
|\Phi^i(0)'| &= |2{u^i}^T u^{i\prime}| = |2{u^i}^T Q^{i\prime} Q^{i^{-1}} u^i| \\
&= |{u^i}^T (Q^{i\prime} Q^{i^{-1}} + Q^{i^{-1}} Q^{i\prime}) u^i| \qquad \text{(since } Q^i, Q^{i\prime} \text{are symmetric matrices)} \\
&= |{u^i}^T Q^{i^{-1}} (Q^i Q^{i\prime} + Q^{i\prime} Q^i) Q^{i^{-1}} u^i| = |{u^i}^T Q^{i^{-1}} ({Q^i}^2)' Q^{i^{-1}} u^i|. \qquad (3.11)
\end{aligned}$$

We let $\nabla\lambda^i := \nabla\lambda^i(\mu, \bar{x})$ and $\lambda^{i\prime} := \frac{\partial \lambda^i(\mu, \bar{x}+td)}{\partial t}\big|_{t=0} = \nabla\lambda_i d$. Note that from (3.4) we have

$$\begin{aligned}
({Q^i}^2)' &= (\Lambda^i \otimes S^{i^{-1}})' = \mu^{-1}(\Lambda^i \otimes \Lambda^i)' = \mu^{-1}(\Lambda^i \otimes \Lambda^{i\prime} + \Lambda^{i\prime} \otimes \Lambda^i) \\
&\qquad \text{(since } \Lambda^{i\prime} = \mathbf{mat}(\nabla\lambda^i d)) \\
&= \mu^{-1}(\Lambda^i \otimes \mathbf{mat}(\nabla\lambda^i d) + \mathbf{mat}(\nabla\lambda^i d) \otimes \Lambda^i). \qquad (3.12)
\end{aligned}$$

Combining (3.11), (3.12), and using (3.4), we obtain

$$\begin{aligned}
|\Phi^i(0)'| &= |{u^i}^T (\Lambda^{i^{-1/2}} \otimes \Lambda^{i^{-1/2}}) \\
&\qquad [(\Lambda^i \otimes \mathbf{mat}(\nabla\lambda^i d) + \mathbf{mat}(\nabla\lambda^i d) \otimes \Lambda^i)](\Lambda^{i^{-1/2}} \otimes \Lambda^{i^{-1/2}}) u^i| \\
&= |{u^i}^T [I \otimes (\Lambda^{i^{-1/2}} \mathbf{mat}(\nabla\lambda^i d) \Lambda^{i^{-1/2}}) + (\Lambda^{i^{-1/2}} \mathbf{mat}(\nabla\lambda^i d) \Lambda^{i^{-1/2}}) \otimes I] u^i| \\
&\le 2\|u^i\|_2^2 \, \|\mathbf{vec}(\Lambda^{i^{-1/2}} \mathbf{mat}(\nabla\lambda^i d) \Lambda^{i^{-1/2}})\|_2 \\
&= 2\|u^i\|_2^2 \, \|(\Lambda^{i^{-1/2}} \otimes \Lambda^{i^{-1/2}})(\nabla\lambda^i d)\|_2 \\
&= 2\mu^{-1/2}\|u^i\|_2^2 \, \|Q^{i^{-1}} \nabla\lambda^i d\|_2 \qquad \text{(noting that } Q^{i^{-1}} = \sqrt{\mu}(\Lambda^{i^{-1/2}} \otimes \Lambda^{i^{-1/2}})) \\
&= 2\mu^{-1/2}\|u^i\|_2^3 \qquad \text{(noting that } Q^{i^{-1}} \nabla\lambda^i d = u^i) \\
&= 2\mu^{-1/2}|\Phi^i(0)|^{3/2} \qquad \text{(since } |\Phi^i(0)| = \|u^i\|_2^2). \qquad (3.13)
\end{aligned}$$

□

We have the following corollary.

COROLLARY 3.1. *The recourse function $\rho(\mu, x)$ is a μ-self-concordant barrier on $\mathcal{F}_1$, and the first stage objective function $\eta(\mu, x) := c^T x + \rho(\mu, x) + \mu \ln\det S$ is a strongly μ-self-concordant barrier on $\mathcal{F}_0$.*

Proof. It is easy to verify that $\mu \ln\det S$ is a strongly μ-self-concordant barrier on $\{x | Ax + s = b, s \in \mathcal{K}^p\}$. The corollary follows from Proposition 2.1.1(ii) in [8]. □

3.3. Parameters of the self-concordant family. The self-concordant family with appropriate parameters is defined in Nesterov and Nemirovskii [8]. They showed that, given such a family, the parameters defining the family allow us to relate the rate at which the barrier parameter μ is varied and the number of Newton steps required to maintain the proximity to the central path. Below is the definition of a strongly self-concordant family adapted to the current setting from the original definition in Nesterov and Nemirovskii [8]. These conditions might look rather technical; nevertheless they simplify our convergence analysis and the accompanying proofs in what follows and explicitly reveal some essential properties of the log-barrier recourse function $\rho(\mu, x)$. They allow us to invoke the interior point convergence theory developed by Nesterov and Nemirovskii [8].

DEFINITION 3.2. *The family of functions* $\{\eta(\mu,\cdot) : \mu > 0\}$ *is strongly self-concordant on* $\mathcal{F}_0$ *with parameter functions* $\alpha(\mu)$, $\gamma(\mu)$, $\nu(\mu)$, $\xi(\mu)$, *and* $\sigma(\mu)$ *if the following six conditions are satisfied:*

C1. $\eta(\mu, x)$ *is concave in* x, *continuous in* $(\mu, x) \in \mathbb{R}_{++} \times \mathcal{F}_0$, *and has three derivatives in* x, *continuous in* $(\mu, x) \in \mathbb{R}_{++} \times \mathcal{F}_0$.

C2. $\nabla\eta(\mu, x)$ *and* $\nabla^2\eta(\mu, x)$ *are continuously differentiable in* μ.

C3. *For any* $\mu \in \mathbb{R}_{++}$, $\eta(\mu, x)$ *is strongly* $\alpha(\mu)$*-self-concordant on* $\mathcal{F}_0$.

C4. *The parameter functions* $\alpha(\mu)$, $\gamma(\mu)$, $\xi(\mu)$, *and* $\sigma(\mu)$ *are continuous positive scalar functions on* $\mu \in \mathbb{R}_{++}$.

C5. *For every* $(\mu, x) \in \mathbb{R}_{++} \times \mathcal{F}_0$ *and* $h \in \mathbb{R}^n$,

$$|\{\nabla\eta(\mu,x)h\}' - \{\ln\nu(\mu)\}'\{\nabla\eta(\mu,x)h\}| \le \xi(\mu)\alpha(\mu)^{1/2}(-h^T\nabla^2\eta(\mu,x)h)^{1/2}.$$

C6. *For every* $(\mu, x) \in \mathbb{R}_{++} \times \mathcal{F}_0$ *and* $h \in \mathbb{R}^n$,

$$|\{h^T\nabla^2\eta(\mu,x)h\}' - \{\ln\gamma(\mu)\}'h^T\nabla^2\eta(\mu,x)h| \le -2\sigma(\mu)h^T\nabla^2\eta(\mu,x)h.$$

We refer the reader to Nesterov and Nemirovskii [8] for the original definition of self-concordant families and their properties. The essence of the above definition is in conditions C5 and C6.

THEOREM 3.1. *The family of functions* $\eta : \mathbb{R}_{++} \times \mathcal{F} \mapsto \mathbb{R}$ *is a strongly self-concordant family with parameters* $\alpha(\mu) = \mu$, $\gamma(\mu) = \nu(\mu) = 1$, $\xi(\mu) = \frac{\sqrt{p+Kr}}{\mu}$, *and* $\sigma(\mu) = \frac{\sqrt{r}}{2\mu}$.

Proof. It is easy to verify that conditions C1–C4 of Definition 3.2 hold. Lemmas 3.2 and 3.3 below show that C5 and C6 are satisfied. □

In Lemmas 3.2 and 3.3 we bound the changes of $\nabla\eta(\mu, x)$ and $\nabla^2\eta(\mu, x)$ as the barrier parameter μ changes. This requires us to calculate the derivatives of $(y^i(\mu, x), \lambda^i(\mu, x), s^i(\mu, x))$ with respect to μ. This derivative is represented by $(y^{i\prime}, \lambda^{i\prime}, s^{i\prime})$. Differentiating (2.9) with respect to μ, we get

$$\begin{aligned} W^{iT}\lambda^{i\prime} &= 0, \\ W^i y^{i\prime} + s^{i\prime} &= 0, \\ (I \otimes S^i)\lambda^{i\prime} + (\Lambda^i \otimes I)s^{i\prime} &= \mathbf{vec}(I). \end{aligned} \tag{3.14}$$

Solving (3.14), we obtain

$$\begin{aligned} y^{i\prime} &= -R^{i-1}W^{iT}s^{i-1}, \\ \lambda^{i\prime} &= \frac{1}{\sqrt{\mu}}Q^iP^i\mathbf{vec}(I), \\ s^{i\prime} &= W^iR^{i-1}W^{iT}s^{i-1}. \end{aligned} \tag{3.15}$$

LEMMA 3.2. *For any $\mu > 0$, $x \in \mathcal{F}_0$, and $h \in \mathbb{R}^n$ we have*

$$|\{\nabla\eta(\mu,x)^T h\}'| \le \left[\frac{-(p+Kr)}{\mu} h^T \nabla^2\eta(\mu,x)^T h\right]^{1/2}.$$

Proof. Differentiating (3.7) with respect to μ and applying (3.15), we get

$$\begin{aligned}\{\nabla\eta(\mu,x)\}' &= -\frac{1}{\sqrt{\mu}}\sum_{i=1}^{K} {T^i}^T Q^i P^i \mathbf{vec}(I) - A^T s^{-1} \\ &= -\frac{1}{\sqrt{\mu}}\sum_{i=1}^{K} {T^i}^T Q^i P^i \mathbf{vec}(I) - A^T(S^{-1/2}\otimes S^{-1/2})\mathbf{vec}(I).\end{aligned}$$

We define

$$B := \left[\frac{1}{\sqrt{\mu}}{T^1}^T Q^1 P^1,\ \ldots,\ \frac{1}{\sqrt{\mu}}{T^K}^T Q^K P^K,\ A^T(S^{-1/2}\otimes S^{-1/2})\right],$$

and let z be a (p^2+Kr^2)-dimensional vector defined by $z := [\mathbf{vec}(I_r),\ldots,\ \mathbf{vec}(I_r), \mathbf{vec}(I_p)]$. We can write

$$\{\nabla\eta(\mu,x)\}' = -Bz. \tag{3.16}$$

Note that $BB^T = \frac{1}{\mu}\sum_{i=1}^{K} {T^i}^T Q^i P^i Q^i T^i + A^T(S^{-1}\otimes S^{-1})A = -\frac{1}{\mu}\nabla^2\eta(\mu,x)$. Now we have

$$-\{\nabla\eta(\mu,x)^T\}'[\nabla^2\eta(\mu,x)]^{-1}\{\nabla\eta(\mu,x)\}' = \frac{1}{\mu}z^T B^T[BB^T]^{-1}Bz \le \frac{1}{\mu}z^T z = \frac{1}{\mu}(p+Kr). \tag{3.17}$$

Now by using norm inequalities and (3.17), it follows that

$$\begin{aligned}|\{\nabla\eta(\mu,x)^T h\}'| &\le \left[-\{\nabla\eta(\mu,x)^T\}'[\nabla^2\eta(\mu,x)]^{-1}\{\nabla\eta(\mu,x)\}'\right]^{1/2}\left[-h^T\nabla^2\eta(\mu,x)h\right]^{1/2} \\ &\le \left[\frac{-(p+Kr)}{\mu}h^T\nabla^2\eta(\mu,x)h\right]^{1/2}. \qquad \square\end{aligned}$$

LEMMA 3.3. *For any $\mu > 0$, $x \in \mathcal{F}^0$ and $h \in \mathbb{R}^n$ we have*

$$|\{h^T\nabla^2\eta(\mu,x)h\}'| \le -\frac{\sqrt{r}}{\mu}h^T\nabla^2\eta(\mu,x)h.$$

Proof. We fix $h \in \mathbb{R}^n$ and let $(\lambda^i, P^i, s^i, Q^i, R^i) := (\lambda^i(\mu,x), P^i(\mu,x), s^i(\mu,x), Q^i(\mu,x), R^i(\mu,x))$. Let us define $u^i := P^i Q^i T^i h$. We have

$$h^T\nabla^2\eta(\mu,x)h = -\sum_{i=1}^{K}{u^i}^T u^i - \mu h^T A^T(S^{-1}\otimes S^{-1})Ah.$$

Following the steps in Lemma 3.1 leading up to (3.11) and the definition of $\eta(\mu,x)$, we have

$$\{h^T\nabla^2\eta(\mu,x)h\}' = -\sum_{i=1}^{K}{u^i}^T {Q^i}^{-1}({Q^i}^2)'{Q^i}^{-1}u^i - h^T A^T(S^{-1}\otimes S^{-1})Ah. \tag{3.18}$$

From (3.15), the definition of Q^i from (3.4), $S^i\Lambda^i = \mu I$, and using $S^i\Lambda^{i\prime} + \Lambda^i S^{i\prime} = I$, $\Lambda^{i^{-1}} = \mu^{-1}S^i$, it follows that

$$\begin{aligned}
&u^{i^T}Q^{i^{-1}}(Q^{i^2})'Q^{i^{-1}}u^i \\
&\quad = u^{i^T}(I \otimes \Lambda^{i^{-1/2}}\Lambda^{i\prime}\Lambda^{i^{-1/2}} - S^{i^{-1/2}}S^{i\prime}S^{i^{-1/2}} \otimes I)u^i \\
&\quad = \mu^{-1}u^{i^T}(I \otimes I - \mu(S^{i^{-1/2}}S^{i\prime}S^{i^{-1/2}} \otimes I + I \otimes S^{i^{-1/2}}S^{i\prime}S^{i^{-1/2}}))u^i \\
&\quad \le \frac{\|u^i\|_2^2}{\mu}\|\mathbf{vec}(I) - 2\mu(S^{i^{-1/2}} \otimes S^{i^{-1/2}})s^{i\prime}\|_2 \\
&\quad = \frac{\|u^i\|_2^2}{\mu}\|\mathbf{vec}(I) - 2\mu(S^{i^{-1/2}} \otimes S^{i^{-1/2}})W^iR^{i^{-1}}W^{i^T}s^{i^{-1}}\|_2 \\
&\quad = \frac{\|u^i\|_2^2}{\mu}\|(I - 2P^i)\mathbf{vec}(I)\|_2 \\
&\quad \le \frac{\sqrt{r}}{\mu}\|u^i\|_2^2 \qquad (\text{since } I - 2P \preceq I, \|(I - 2P^i)\|_2 \le 1). \qquad (3.19)
\end{aligned}$$

From (3.18) and (3.19), we obtain for any $h \in R^n$

$$|\{h^T\nabla^2\eta(\mu,x)h\}'| \le \frac{\sqrt{r}}{\mu}\sum_{i=1}^K u^{i^T}u^i + h^TA^T(S^{-1}\otimes S^{-1})Ah = -\frac{\sqrt{r}}{\mu}h^T\nabla^2\eta(\mu,x)h. \qquad \square$$

4. Interior point decomposition algorithms for TSSDP. Once it is established that the family of functions $\{\eta(\mu,\cdot) : \mu > 0\}$ is strongly self-concordant, the development of primal path following interior point methods is straightforward. These methods reduce μ by a factor at each iteration and seek to approximate the minimizer $x(\mu)$ for each μ by taking one or more Newton steps. The novelty of the algorithm in the context of TSSDP is in computing the Newton direction from the solutions of the decomposed second stage problems. As μ varies, the minimizers $x(\mu)$ form the central path. By tracing the central path as $\mu \to 0$, this procedure will generate a strictly feasible ϵ-solution to (2.5).

For a given μ the optimality condition for the problem (2.5) is

$$\nabla\eta(\mu, x(\mu)) = 0. \qquad (4.1)$$

Hence, at a feasible point x the Newton direction is given by

$$\Delta x = -[\nabla^2\eta(\mu,x)]^{-1}\nabla\eta(\mu,x). \qquad (4.2)$$

Note that although problems (2.5)–(2.7) and (2.10) share the same central path, the associated Newton directions are not identical and lead to different ways of path following. A conceptual primal path following algorithm is given below.

THE DECOMPOSITION ALGORITHM. Here $\beta > 0, \gamma \in (0,1)$ and $\theta > 0$ are suitable scalars. We make their values more precise in Theorems 4.1 and 4.2. The desired precision ϵ, an initial point $x^0 \in \mathcal{F}_0$, and μ^0 are given as inputs.

Initialization. $x = x^0$; $\mu = \mu^0$.

Step 1 (Newton or damped Newton iterations).

1.1. For all i solve the optimality conditions (2.9) to find $(y^i(\mu,x), s^i, \lambda^i(\mu,x))$.

1.2. Compute the Newton direction Δx from (4.2).

1.3. Let $\delta(\mu,x)=\sqrt{-\frac{1}{\mu}\Delta x^T\nabla^2\eta(\mu,x)\Delta x}$. If $\delta\le\beta$, go to Step 2.

1.4. Set $x=x+\theta\Delta x$ and go to Step 1.1.

Step 2 (termination check). If $\mu\le\epsilon$, stop; otherwise set $\mu=\gamma\mu$ and go to Step 1.1.

In the above algorithm we assume that we can find exact solutions of the optimality conditions (2.9). This assumption considerably simplifies the complexity analysis. In a practical implementation of this algorithm (such as the one in [7]) we use approximate solutions of the optimality conditions (2.9) to construct the Newton direction (4.2).

Theorems 4.1 and 4.2 give two standard complexity results for the generic primal interior point method. In the short-step version of the algorithm, barrier parameter μ is decreased by a factor $1-\sigma/\sqrt{n+m}$ ($\sigma>0$) in each iteration.

An iteration of the short-step algorithm is performed as follows. At the beginning of iteration k, x^k is close to the central path; i.e., $\delta(\mu^k,x^k)\le\beta$. After reducing the parameter from μ^k to $\mu^{k+1}=\gamma\mu^k$, we will have $\delta(\mu^{k+1},x^k)\le 2\beta$. Then a Newton step with step size $\theta=1$ is taken, resulting in a new point x^{k+1} with $\delta(\mu^{k+1},x^{k+1})\le\beta$. We have the following theorem.

THEOREM 4.1. *Let μ^0 be the initial barrier parameter, $\epsilon>0$ the stopping criterion, and $\beta=(2-\sqrt{3})/2$. If the starting point x^0 is sufficiently close to the central path, i.e., $\delta(\mu^0,x^0)\le\beta$, then the short-step algorithm reduces the barrier parameter μ at a linear rate and terminates within $O(\sqrt{p+Kr}\ln\mu^0/\epsilon)$ executions of Steps* 1.1–1.4 *in Algorithm* 1.

Proof. For the proof, see section 5.1. ∎

In the long-step version we decrease the barrier parameter μ by an arbitrary constant factor $\lambda\in(0,1)$. This has a potential for much faster progress; however, several damped Newton steps might be needed for restoring the proximity to the central path. We have the following theorem.

THEOREM 4.2. *Let μ^0 be the initial barrier parameter, $\epsilon>0$ be the stopping criterion, and $\beta=1/6$. If the starting point x^0 is sufficiently close to the central path, i.e., $\delta(\mu^0,x^0)\le\beta$, then the long-step algorithm reduces the barrier parameter μ at a linear rate and terminates within $O((p+Kr)\ln\mu^0/\epsilon)$ executions of Steps* 1.1–1.4 *in Algorithm* 1.

Proof. For the proof, see section 5.2. ∎

5. Convergence proof for short- and long-step algorithms. Part (i) of the following proposition follows directly from the definition of self-concordance and is due to Nesterov and Nemirovskii [8, Theorem 2.1.1]. Part (ii) is a corollary of part (i) and is given in Zhao [15] without a proof.

PROPOSITION 5.1. *For any $\mu>0$, $x\in\mathcal{F}^0$, and Δx computed from* (4.2), *let $\delta:=\sqrt{-\frac{1}{\mu}\Delta x^T\nabla^2\eta(\mu,x)\Delta x}$. Then, for $\delta<1$, $\tau\in[0,1]$, and any $h\in\mathbb{R}^n$ we have*

$$\text{(i)}\quad -(1-\tau\delta)^2h^T\nabla^2\eta(\mu,x)h\le -h^T\nabla^2\eta(\mu,x+\tau\Delta x)h\le -(1-\tau\delta)^{-2}h^T\nabla^2\eta(\mu,x)h,$$

$$\text{(ii)}\quad |h_1^T[\nabla^2\eta(\mu,x+\tau\Delta x)-\nabla^2\eta(\mu,x)]h_2|\le[(1-\tau\delta)^{-2}-1]\sqrt{-h_1^T\nabla^2\eta(\mu,x)h_1}\sqrt{-h_2^T\nabla^2\eta(\mu,x)h_2}.$$

For the estimation of the number of Newton steps needed for recentering we use two different merit functions to measure the progress of Newton iterates. We use $\delta(\mu,x)$ for the short-step algorithm and the first stage objective $\eta(\mu,x)$ for the long-

step algorithm. The following lemma is due to Theorem 2.2.3 in [8] and describes the behavior of the Newton method as applied to $\eta(\mu, \cdot)$.

LEMMA 5.1. *Let $\mu > 0$ and $x \in \mathcal{F}^0$, Δx be the Newton direction calculated at x from (4.2); $\delta := \delta(\mu, x) = \sqrt{-\frac{1}{\mu}\Delta x^T \nabla^2 \eta(\mu, x)\Delta x}$, $x^+ = x + \Delta x$, Δx^+ be the Newton direction calculated at x^+; and $\delta(\mu, x^+) := \sqrt{-\frac{1}{\mu}\Delta x^{+T} \nabla^2 \eta(\mu, x^+)\Delta x^+}$. Then,*

(i) *if $\delta < 2 - \sqrt{3}$, then*

$$\delta(\mu, x^+) \leq \left(\frac{\delta}{1-\delta}\right)^2 \leq \frac{\delta}{2};$$

(ii) *if $\delta \geq 2 - \sqrt{3}$, then*

$$\eta(\mu, x) - \eta(\mu, x + \overline{\theta}\Delta x) \geq \mu[\delta - \ln(1+\delta)],$$

for $\overline{\theta} = (1+\delta)^{-1}$.

5.1. Complexity of the short-step algorithm. We now show that in this version of the algorithm a single Newton step is sufficient for recentering after updating the barrier parameter μ. To this end we make use of Theorem 3.1.1 in [8], which is restated for the present context in the next proposition.

PROPOSITION 5.2. *Let $\varphi_\kappa(\eta; \mu, \mu^+) := \left(\frac{1+r}{2} + \frac{\sqrt{p+Kr}}{\kappa}\right) \ln \gamma^{-1}$. Assume that $\delta(\mu, x) < \kappa$ and $\mu^+ := \gamma\mu$ satisfies*

$$\varphi_\kappa(\eta; \mu, \mu^+) \leq 1 - \frac{\delta(\mu, x)}{\kappa}.$$

Then $\delta(\mu^+, x) < \kappa$.

LEMMA 5.2. *Let $\mu^+ = \gamma\mu$, where $\gamma = 1 - \sigma/\sqrt{p+Kr}$ and $\sigma \leq 0.1$. Furthermore let $\beta = (2-\sqrt{3})/2$. If $\delta(\mu, x) \leq \beta$, then $\delta(\mu^+, x) \leq 2\beta$.*

Proof. Let $\kappa = 2\beta = 2 - \sqrt{3}$. It is easy to verify that with $\sigma \leq 0.1$, μ^+ satisfies

$$\begin{aligned}\varphi_\kappa(\eta; \mu, \mu^+) &= \left(\frac{1+r}{2} + \frac{\sqrt{p+Kr}}{\kappa}\right) \ln(1 - \sigma/\sqrt{p+Kr})^{-1} \\ &\leq \frac{1}{2} \leq 1 - \frac{\delta(\mu, x)}{\kappa}.\end{aligned}$$

Now Proposition 5.2 implies

$$\delta(\mu^+, x) \leq \kappa = 2\beta. \qquad \square$$

From Lemmas 5.1 and 5.2 it is clear that we can reduce μ by the factor $\gamma = 1 - \sigma/\sqrt{p+Kr}$, $\sigma < 0.1$, at each iteration and that a single Newton step is sufficient to restore proximity to the central path. Hence, Theorem 4.1 follows.

5.2. Complexity of the long-step algorithm. For the analysis of the long-step algorithm we use η as the merit function since the iterates generated by the less conservative long-step algorithm may violate the condition, $\delta < 2 - \sqrt{3}$, required in part (i) of Lemma 5.1. Our analysis follows the steps in Zhao [15].

Assume that we have a point x^{k-1} sufficiently close to $x(\mu^{k-1})$. Next we reduce the barrier parameter from μ^{k-1} to $\mu^k = \gamma\mu^{k-1}$, where $\gamma \in (0, 1)$. While searching for a point x^k that is sufficiently close to $x(\mu^k)$, the long-step algorithm generates a

finite sequence of points (inner iterates) $p^1, \ldots, p^N \in \mathcal{F}_0$, and we finally set $x^k = p^N$. We need to determine an upper bound on N, the number of Newton iterations needed for recentering. Let

$$\phi(\mu, x) := \eta(\mu, x(\mu)) - \eta(\mu, x).$$

The next lemma gives upper bounds on $\phi(\mu^{k-1}, x)$ and $\phi'(\mu^{k-1}, x)$, respectively, for any $\mu > 0$ and $x \in \mathcal{F}_0$. They facilitate us bounding $\phi(\mu^k, x)$.

LEMMA 5.3. *Let $\mu > 0$ and $x \in \mathcal{F}_0$. We denote $\tilde{\Delta}x := x(\mu) - x$ and define*

$$\tilde{\delta}(\mu, x) := \sqrt{-\frac{1}{\mu}\tilde{\Delta}x^T \nabla^2 \eta(\mu, x)\tilde{\Delta}x}.$$

For any $\mu > 0$ and $x \in \mathcal{F}_0$, if $\tilde{\delta} < 1$, then

$$\phi(\mu, x) \le \mu \left[\frac{\tilde{\delta}}{1-\tilde{\delta}} + \ln(1-\tilde{\delta})\right], \tag{5.1}$$

$$|\phi'(\mu, x)| \le -\sqrt{p + Kr}\ln(1-\tilde{\delta}). \tag{5.2}$$

Proof.

$$\phi(\mu, x) = \eta(\mu, x(\mu)) - \eta(\mu, x) = \int_0^1 \nabla\eta(\mu, x + \tau\tilde{\Delta}x)^T \tilde{\Delta}x d\tau.$$

Since $x(\mu)$ is the optimal solution of (2.5), it satisfies the optimality conditions (4.1). Hence,

$$\begin{aligned}
\phi(\mu, x) &= \int_0^1 \int_0^\tau \tilde{\Delta}x^T \nabla^2 \eta(\mu, x + \alpha\tilde{\Delta}x)\tilde{\Delta}x d\alpha d\tau \\
&\le \int_0^1 \int_0^\tau \frac{\mu\tilde{\delta}^2}{(1-\alpha\tilde{\delta})^2} d\alpha d\tau \qquad \text{(using Proposition 5.1 (i))} \\
&= \mu \left[\frac{\tilde{\delta}}{1-\tilde{\delta}} + \ln(1-\tilde{\delta})\right]. \qquad (5.3)
\end{aligned}$$

This proves (5.1). Now, for any $\mu > 0$, by applying the chain rule and using (4.1) we have

$$\begin{aligned}
\phi'(\mu, x) &= \eta'(\mu, x(\mu)) - \eta'(\mu, x) + \nabla\eta(\mu, x(\mu))^T x'(\mu) \\
&= \eta'(\mu, x(\mu)) - \eta'(\mu, x). \qquad (5.4)
\end{aligned}$$

From (5.4), applying the mean-value theorem, we obtain

$$\begin{aligned}
|\phi'(\mu, x)| &= \left| \int_0^1 \{\nabla\eta(\mu, x + \tau\tilde{\Delta}x)^T\}' \tilde{\Delta}x \, d\tau \right| \\
&\le \int_0^1 \left[-\tilde{\Delta}x^T \nabla^2 \eta(\mu, x + \tau\tilde{\Delta}x)\tilde{\Delta}x\right]^{1/2} \\
&\quad \left[-\{\nabla\eta(\mu, x + \tau\tilde{\Delta}x)^T\}'[\nabla^2\eta(\mu, x + \tau\tilde{\Delta}x)]^{-1}\{\nabla\eta(\mu, x + \tau\tilde{\Delta}x)^T\}'\right]^{1/2} d\tau. \qquad (5.5)
\end{aligned}$$

From (5.5), and using (3.17) from the proof of Lemma 3.2 and Proposition 5.1(i), we get

$$|\phi'(\mu,x)| \le \int_0^1 \frac{\sqrt{-\tilde{\Delta}x^T\nabla^2\eta(\mu,x)\tilde{\Delta}x}}{1-\tilde{\delta}+\tau\tilde{\delta}}\sqrt{\frac{p+Kr}{\mu}}\,d\tau$$
$$\le \int_0^1 \frac{\sqrt{\mu}\tilde{\delta}}{1-\tilde{\delta}+\tau\tilde{\delta}}\sqrt{\frac{p+Kr}{\mu}}\,d\tau = -\sqrt{p+Kr}\ln(1-\tilde{\delta}). \qquad \square$$

LEMMA 5.4. *Let $\mu > 0$ and $x \in \mathcal{F}_0$ be such that $\tilde{\delta} < 1$, where $\tilde{\delta}$ is defined in Lemma 5.3. Let $\mu^+ = \gamma\mu$ with $\gamma \in (0,1)$. Then,*

$$\eta(\mu^+, x(\mu^+)) - \eta(\mu^+, x) \le O(p+Kr)\mu^+.$$

Proof. By differentiating (5.4) with respect to μ, we obtain

$$\phi''(\mu,x) = \eta''(\mu,x(\mu)) + \nabla\eta'(\mu,x(\mu))^T x'(\mu) - \eta''(\mu,x). \tag{5.6}$$

Now we will bound the two terms on the right-hand side of (5.6) separately. From the definition of $\eta(\mu,x)$ in (2.5) we see that for $\mu > 0$ and $x \in \mathcal{F}_0$, $\eta''(\mu,x) = \sum_{i=1}^K \rho^{i\prime\prime}(\mu,x)$. Differentiating $\rho^i(\mu,x)$ and using (3.15), we obtain

$$\begin{aligned}\rho^{i\prime}(\mu,x) &= d^{i^T}y^{i\prime} + \ln\det S^i + \mu s^{i^{-T}}s^{i\prime}\\ &= \ln\det S^i + (-d^i + W^{i^T}\lambda^i)^T R^{i^{-1}} W^{i^T} s^{i^{-1}}\\ &= \ln\det S^i. \end{aligned} \tag{5.7}$$

Differentiating (5.7) once more and using (3.15) for $\mu > 0$ and $x \in \mathcal{F}_0$, we get

$$\rho^{i\prime\prime}(\mu,x) = s^{i^{-T}}s^{i\prime} = s^{i^{-T}}W^i R^{i^{-1}} W^{i^T} s^{i^{-1}} = s^{i^{-T}}Q^{i^{-1}}P^i Q^{i^{-1}} s^{i^{-1}}.$$

Since P^i in (3.5) is an orthogonal projection matrix $\rho^{i\prime\prime}(\mu,x) \ge 0$, hence $\eta(\mu,x)$ is a convex function in μ. We also have

$$\rho^{i\prime\prime}(\mu,x(\mu)) \le s^{i^{-T}}Q^{i^{-2}}s^{i^{-1}} = s^{i^{-T}}\lambda^{i^{-1}} = \frac{r}{\mu}$$

and thus

$$\eta''(\mu,x(\mu)) \le \frac{Kr}{\mu}. \tag{5.8}$$

Differentiating the optimality condition of the first stage problem (2.5), we observe

$$x(\mu)' = -[\nabla^2\eta(\mu,x(\mu))]^{-1}\nabla\eta'(\mu,x(\mu)). \tag{5.9}$$

Hence, we have

$$\begin{aligned}\nabla\eta'(\mu,x(\mu))^T x'(\mu) &= -\nabla\eta'(\mu,x(\mu))^T[\nabla^2\eta(\mu,x(\mu))]^{-1}\nabla\eta'(\mu,x(\mu))\\ &\le \mu^{-1}(p+Kr).\end{aligned} \tag{5.10}$$

In the last inequality we used (3.17), which is valid for any $\mu > 0$ and $x \in \mathcal{F}_0$. Combining (5.8), (5.10), and using $\eta''(\mu,x) > 0$, we have

$$\phi''(\mu,x) \le \mu^{-1}(p+2Kr). \tag{5.11}$$

Now in view of Lemma 5.3 and (5.11), we have

$$\begin{aligned}\phi(\mu^+,x) &= \phi(\mu,x)+\phi'(\mu,x)(\mu^+-\mu)+\int_\mu^{\mu^+}\int_\mu^\tau \phi''(\nu,x)\ d\nu\ d\tau\\ &\le \mu\left[\frac{\tilde\delta}{1-\tilde\delta}+\ln(1-\tilde\delta)\right]-\sqrt{p+Kr}\ln(1-\tilde\delta)(\mu-\mu^+)\\ &\quad +(p+2Kr)\int_\mu^{\mu^+}\int_\mu^\tau \nu^{-1}\ d\nu\ d\tau\\ &\le \mu\left[\frac{\tilde\delta}{1-\tilde\delta}+\ln(1-\tilde\delta)\right]-\sqrt{p+Kr}\ln(1-\tilde\delta)(\mu-\mu^+)\\ &\quad +(p+2Kr)\ln\gamma^{-1}\ (\mu-\mu^+).\end{aligned}\tag{5.12}$$

Since γ and $\tilde\delta$ are absolute constants, and given the fact that $\eta(\mu,x)$ is a strictly convex function in μ (implying $\eta''(\mu,x)>0$), we have a proof of this lemma. ∎

Note that Lemmas 5.3 and 5.4 require $\tilde\delta$ to be less than one. However, we cannot evaluate $\tilde\delta$ since we do not explicitly know the points $x(\mu)$ forming the central path. Nonetheless we can evaluate δ and $\tilde\delta$ in proportion to δ, as shown in the following lemma.

LEMMA 5.5. *For any given $\mu>0$, $x\in\mathcal{F}_0$, let Δx be the Newton direction defined in (4.2) and $\tilde\Delta x := x-x(\mu)$. We denote*

$$\delta:=\delta(\mu,x)=\sqrt{-\frac{1}{\mu}\Delta x^T\nabla^2\eta(\mu,x)\Delta x}\quad\text{and}\quad \tilde\delta:=\tilde\delta(\mu,x)=\sqrt{-\frac{1}{\mu}\tilde\Delta x^T\nabla^2\eta(\mu,x)\tilde\Delta x}.$$

If $\delta\le 1/6$, then

$$\frac{2}{3}\delta\le\tilde\delta\le 2\delta.\tag{5.13}$$

Proof. Let $H:=\nabla^2\eta(\mu,x)$, $g:=\nabla\eta(\mu,x)$. We denote $\overline{g}:=g+H\tilde\Delta x$. Note that

$$\tilde\Delta x=-\Delta x+H^{-1}\overline{g}.\tag{5.14}$$

By applying the triangle inequality to (5.14), we obtain

$$\tilde\delta\le\delta+\sqrt{-\frac{1}{\mu}\overline{g}^TH^{-1}\overline{g}}.\tag{5.15}$$

It is straightforward to verify that

$$-\overline{g}H^{-1}\overline{g}=\max\{h^THh-2h^T\overline{g}\mid h\in\mathbb{R}^n\}.$$

Now in consideration of Proposition 5.1(ii), we have

$$\begin{aligned}-h^T\overline{g} &= -\int_0^1 h^T[\nabla^2\eta(\mu,x)-\nabla^2\eta(\mu,x-(1-\tau)\tilde\Delta x)]\tilde\Delta x d\tau\\ &\le \int_0^1[(1-(1-\tau)\tilde\delta)^{-2}-1]d\tau\ \sqrt{-\tilde\Delta x^TH\tilde\Delta x}\sqrt{-h^THh}\\ &= \frac{\sqrt{\mu}\tilde\delta^2}{1-\tilde\delta}\sqrt{-h^THh}.\end{aligned}$$

Hence,

$$-\overline{g}H^{-1}\overline{g} \le \max\left\{ h^T Hh + \frac{2\sqrt{\mu}\tilde{\delta}^2}{1-\tilde{\delta}}\sqrt{-h^T Hh} \mid h \in \mathbb{R}^n \right\}$$

$$= \frac{\mu\tilde{\delta}^4}{(1-\tilde{\delta})^2}. \tag{5.16}$$

Combining (5.15) and (5.16), we obtain

$$\tilde{\delta} \le \delta + \frac{\tilde{\delta}^2}{(1-\tilde{\delta})}. \tag{5.17}$$

When $\delta \le \frac{1}{6}$, the quadratic inequality (5.17) implies $\tilde{\delta} \le 2\delta$. The condition $\delta \le \frac{1}{6}$ is eventually reached, since the inner iterations of the long-step algorithm converge.

From (5.15), exchanging positions of Δx and $\tilde{\Delta} x$ and following the above steps, we obtain $\delta \le \tilde{\delta} + \frac{\tilde{\delta}^2}{(1-\delta)}$, which in turn implies $\delta \le \frac{3}{2}\tilde{\delta}$, since $\tilde{\delta} \le 2\delta \le \frac{1}{3}$. ∎

Lemma 5.1 implies that each inner iteration decreases the value of η by at least $\mu[\delta - \ln(1+\delta)]$. Therefore, in view of Lemmas 5.1 and 5.4 after reducing μ by a factor $\gamma \in (0,1)$, at most $O(p + Kr)$ Newton iterations are needed for recentering. In the long-step version of our algorithm we need to update the barrier parameter μ no more than $O(\ln \mu^0/\epsilon)$ times.

Theorem 4.2 follows from Lemma 5.1(ii), Lemma 5.4, and Lemma 5.5.

6. Concluding remarks. We have described and analyzed short- and long-step interior point decomposition algorithms that follow the primal central trajectory of the first stage problem. At each iteration, these algorithms generate gradient and Hessian information for the first stage problem using optimal dual solutions of the second stage barrier problems and then update the primal first stage solution taking a step along the Newton direction. Although this framework is attractive from the decomposition point of view, it has several limitations that need further work to develop practical implementations. These include (i) generating a good starting point, (ii) development of a practical first stage step length selection procedure, (iii) a practical strategy for reducing μ in our context, (iv) adaptive addition of scenarios, (v) computation of approximate solutions of the second stage problems, and (vi) a proper choice of ϵ to terminate the algorithm. These issues are addressed in a companion computational paper [7].

Acknowledgement. We would like to thank two anonymous referees, whose careful reading of the first draft of this paper resulted in several improvements and corrections.

REFERENCES

[1] O. Bahn, O. du Merle, J.-L. Goffin, and J. P. Vial, *A cutting plane method from analytic centers for stochastic programming*, Math. Programming, 69 (1995), pp. 45–73.

[2] J. R. Birge, C. J. Donohue, D. F. Holmes, and O. G. Svintsiski, *A parallel implementation of the nested decomposition algorithm for multistage stochastic linear programs*, Math. Programming, 75 (1996), pp. 327–352.

[3] J. R. Birge and D. F. Holmes, *Computing block-angular Karmarkar projections with applications to stochastic programming*, Comput. Optim. Appl., 1 (1992), pp. 245–276.

[4] T. Carpenter, I. Lustig, and J. Mulvey, *Formulating stochastic programs for interior point methods*, Oper. Res., 39 (1991), pp. 757–770.

[5] J. Czyzyk, R. Fourer, and S. Mehrotra, *A study of the augmented system and column splitting approaches for solving two-stage stochastic linear programs by interior point methods*, ORSA J. Comput., 7 (1995), pp. 474–490.

[6] S. Mehrotra and M.G. Özevin, *Decomposition-based interior point methods for two-stage stochastic convex quadratic programs with recourse*, Oper. Res., to appear.

[7] S. Mehrotra and M.G. Özevin, *On the Implementation of Interior Decomposition Algorithms for Stochastic Conic Programs*, Technical Report 2005-04, Industrial Engineering and Management Sciences Department, Northwestern University, Evanston, IL, 2005.

[8] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM Stud. Appl. Math. 13, SIAM, Philadelphia, 1994.

[9] M. V. Ramana, L. Tunçel, and H. Wolkowicz, *Strong duality for semidefinite programming*, SIAM J. Optim., 7 (1997), pp. 641–662.

[10] J Renegar, *A Mathematical View of Interior-Point Methods in Convex Optimization*, MPS/SIAM Ser. Optim. 3, SIAM, Philadelphia, 2001.

[11] R. T. Rockafellar and R. J.-B. Wets, *A Lagrangian finite generation technique for solving linear-quadratic problems in stochastic programming*, Math. Program. Study, 28 (1986), pp. 63–93.

[12] R. T. Rockafellar and R. J.-B. Wets, *Scenarios and policy aggregation in optimization under uncertainty*, Math. Oper. Res., 16 (1991), pp. 119–147.

[13] R. M. Van Slyke and R. Wets, *L-shaped linear programs with applications to optimal control and stochastic linear programming*, SIAM J. Appl. Math., 17 (1969), pp. 638–663.

[14] H. Wolkowicz, R. Saigal, and L. Vandenberghe, eds., *Handbook of Semidefinite Programming: Theory, Algorithms,and Applications*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000.

[15] G. Zhao, *A log-barrier method with Benders decomposition for solving two-stage stochastic linear programs*, Math. Program., 90 (2001), pp. 507–536.

SIAM J. OPTIM.
Vol. 18, No. 1, pp. 223–241

A COPOSITIVE PROGRAMMING APPROACH TO GRAPH PARTITIONING*

JANEZ POVH† AND FRANZ RENDL‡

Abstract. We consider 3-partitioning the vertices of a graph into sets S_1, S_2, and S_3 of specified cardinalities, such that the total weight of all edges joining S_1 and S_2 is minimized. This problem is closely related to several NP-hard problems like determining the bandwidth or finding a vertex separator in a graph. We show that this problem can be formulated as a linear program over the cone of completely positive matrices, leading in a natural way to semidefinite relaxations of the problem. We show in particular that the spectral relaxation introduced by Helmberg et al. (1995) can equivalently be formulated as a semidefinite program. Finally we propose a tightened version of this semidefinite program and show on some small instances that this new bound is a significant improvement over the spectral bound.

Key words. semidefinite programming, copositive programming, graph partitioning problem, bandwidth problem, vertex separator problem

AMS subject classifications. 90C22, 90C27

DOI. 10.1137/050637467

1. Introduction.

We consider the following partition problem on graphs, and we denote it as the MIN-CUT problem (MCP). Let $G = (V, E)$ be an undirected graph on n vertices, given by its (weighted) adjacency matrix $A \geq 0$, so $a_{ij} > 0$ implies the edge $(ij) \in E(G)$ with weight a_{ij}. For given integers m_1, m_2, and m_3 summing to n, we are interested in the following NP-complete problem: find subsets S_1, S_2, and S_3 of $V(G)$ with cardinalities m_1, m_2, and m_3, respectively, such that the total weight of edges between S_1 and S_2 is minimal. More formally, let (S_1, S_2, S_3) be a partition of V with $|S_i| = m_i$ for $i = 1, 2, 3$. The total weight of edges between sets S_1 and S_2 will be denoted as $\mathrm{cut}(S_1, S_2)$. Hence

$$\mathrm{cut}(S_1, S_2) = \sum_{i \in S_1,\, j \in S_2} a_{ij}.$$

We define the MCP as the following optimization problem:

$$\text{(MCP)} \qquad \begin{array}{rl} \min & \mathrm{cut}(S_1, S_2) \\ \text{such that} & (S_1, S_2, S_3) \text{ partitions } V(G) \\ & \text{and } |S_i| = m_i,\ i = 1, 2, 3. \end{array}$$

The optimal value of this problem will be denoted as OPT_{MC}.

REMARK 1. *If $m_1 = 0$ or $m_2 = 0$, then the MCP is trivial: $OPT_{MC} = 0$. Therefore, we assume from now on that $1 \leq m_1 \leq m_2$. If $m_3 = 0$, $m_1 = \lfloor \frac{n}{2} \rfloor$, and $m_2 = \lceil \frac{n}{2} \rceil$, we get the NP-complete bisection problem as a special case (see [7]).*

*Received by the editors August 3, 2005; accepted for publication (in revised form) September 15, 2006; published electronically March 19, 2007. Partial support by the Marie Curie Research Training Network MCRTN-CT-2003-504438 (ADONET) is gratefully acknowledged.

http://www.siam.org/journals/siopt/18-1/63746.html

†Faculty of Logistics, University in Maribor, Mariborska cesta 3, Celje, Slovenia (janez.povh@uni-mb.si).

‡Institut für Mathematik, Universität Klagenfurt, Universitätsstraße 65-67, 9020 Klagenfurt, Austria (franz.rendl@uni-klu.ac.at).

The MCP by itself may seem like an artificial optimization problem. It can however serve as a powerful tool to solve some fundamental graph optimization problems. It is connected to the (balanced) vertex separator problem, where the objective is to find a minimal subset of $V(G)$, whose removal disconnects the graph into two subgraphs of roughly equal size. If $OPT_{MC} = 0$, then the graph G underlying A has a *vertex separator* of size m_3 and its connectivity is at most m_3 (see [10] for more details). On the other hand, if $OPT_{MC} > 0$, then the *bandwidth* of the matrix A is at least $m_3 + 1$ (see [10]). The ability to solve (or at least approximate) the MCP therefore has strong impact on these graph problems. The MCP is a special instance of more general graph partitioning problems, where one is interested in a partition of $V(G)$ into k disjoint subsets $S_1, \ldots, S_k$ with cardinalities $m_1 \leq m_2 \leq \cdots \leq m_k$, $\sum_i m_i = |V(G)|$, such that the total weight of edges between some subsets is minimized. A survey on the graph partitioning problem and related problems is given in [13]. Graph partitioning, bandwidth minimization, and vertex separator problems appear in a wide range of applications, from numerical linear algebra to floor planning and analysis of bottlenecks in communication networks. In parallel computing, partitioning the set of tasks among processors in order to minimize the communication between processors is another instance of a graph partitioning problem. A comprehensive survey with results in this area up to 1995 is contained in [1]. Formulating the partitions using vertex variables leads to a quadratic cost function with linear and quadratic constraints in binary variables; see (1)–(5) below. Maintaining the orthogonality condition (2) leads to spectral relaxations based on the Hoffman–Wielandt inequality; see [10, 17]. In [10, 17], these relaxations are investigated for the MCP. The quality of this approach has also been studied in [8]. The spectral relaxation of the MCP from [10] is attractive because of the closed form optimal solution; see (15) in section 3 below. The drawback of this model however is that further refinements, like adding sign constraints, make it intractable. It is the purpose of the present paper to overcome these difficulties, and extend and strengthen the spectral approach. Here are our main contributions.

(i) We first formulate the MCP as a linear program over the cone of completely positive matrices; see section 2. This does not make the problem tractable, since linear optimization over this cone is NP-hard [14], but suggests a new family of tractable relaxations, which we get by approximating the copositive constraint with a tractable one, for example, by using the hierarchy of cones, suggested by Parrilo [15], which approximates the cone of completely positive matrices arbitrarily close.

(ii) The conic formulation of the MCP leads to various semidefinite programming (SDP) relaxations of increasing complexity and strength. In section 3 we show that the spectral model from [10] corresponds to a specific semidefinite program, obtained by approximating the cone of completely positive matrices by the cone of positive semidefinite matrices. The proof of this result is rather involved, and we break it down into several smaller steps. It is given in section 4. As in [10] we provide a closed form solution of this semidefinite program (subsection 4.3).

(iii) Finally, we investigate further tightenings of the SDP relaxation in section 5. This opens the way to powerful new approximations of the bandwidth and vertex separator problems. We provide some preliminary computational results which clearly indicate the potential of the new relaxations.

We point out that similar results have been shown recently for other combinatorial optimization problems. De Klerk and Pasechnik [11] have shown that computing the stability number of a graph is equivalent to solving a copositive program. Anstreicher and Wolkowicz [2] have shown that the spectral relaxation of the quadratic assignment

problem can equivalently be formulated as a semidefinite program. SDP also turned out to be a useful tool to get tractable relaxations for the graph partitioning problem (see [19]) and the vertex separator problem (see [6]). Finally, de Sousa and Balas have recently proposed an integer linear programming approach combined with a branch and cut algorithm to get minimal balanced vertex separators; see [18].

1.1. Notation. We denote the ith standard unit vector by e_i, while the vector of all ones is $u_n \in R^n$ (or u if dimension n is obvious). The square matrix of all ones is J_n (or J) and the identity matrix is $I = (\delta_{ij})$. We set $E_{ij} = e_i e_j^T$ and its symmetrization is $B_{ij} = \frac{1}{2}(E_{ij} + E_{ji})$. In this paper we consider the following sets of matrices. The vector space of real symmetric $n \times n$ matrices is denoted by $\mathcal{S}_n = \{X \in R^{n\times n} \colon X = X^T\}$. The cone of $n \times n$ positive semidefinite matrices is $\mathcal{S}_n^+ = \{X \in \mathcal{S}_n \colon y^T X y \geq 0 \ \forall y \in R^n\}$. The cone of $n \times n$ copositive matrices is denoted by $\mathcal{C}_n = \{X \in \mathcal{S}_n \colon y^T X y \geq 0 \ \forall y \in R_+^n\}$, the cone of $n \times n$ completely positive matrices is $\mathcal{C}_n^* = \{X = \sum_{i=1}^k y_i y_i^T,\ k \geq 1,\ y_i \in R_+^n \ \forall i = 1, \ldots, k\}$, and the cone of $n \times n$ symmetric nonnegative matrices is $\mathcal{N}_n = \{X \in \mathcal{S}_n \colon x_{ij} \geq 0 \ \forall i, j\}$. We also use $X \succeq 0$ for $X \in \mathcal{S}_n^+$ and $X \geq 0$ for an elementwise nonnegative matrix. A linear program over $\mathcal{S}_n^+$ is called a semidefinite program, while a linear program over $\mathcal{C}_n$ or $\mathcal{C}_n^*$ is called a copositive program.

The sign $\otimes$ stands for Kronecker product, while the matrices V_i and W_j denote $V_i = e_i u_3^T \in R^{3\times 3}$, $W_j = e_j u_n^T \in R^{n\times n}$, $1 \leq i \leq 3$, $1 \leq j \leq n$. When we consider a matrix $X \in R^{m\times n}$ as a vector from R^{mn}, we write this vector as $\mathrm{vec}(X)$ or x. The $\langle \cdot, \cdot \rangle$ denote the standard scalar product. For $u, v \in R^n$ we have $\langle u, v \rangle = u^T v$ and for $X, Y \in R^{m\times n}$ we have $\langle X, Y \rangle = \mathrm{trace}(X^T Y)$. For matrix columns and rows we will use MATLAB notation; hence $X(i,:)$ and $X(:,i)$ will stand for the ith row and column, respectively. If $a \in R^n$, then $\mathrm{Diag}(a)$ is an $n \times n$ diagonal matrix with a on the main diagonal. When P is the name of the optimization problem then OPT_P denotes its optimal value.

2. The MCP as a conic linear program. We first use the partition formulation of the MCP to express the MCP as a quadratic program in nonnegative variables. Following [10] we represent partitions (S_1, S_2, S_3) of $V(G)$ by $n \times 3$ matrices X, where

$$x_{ij} = \begin{cases} 1 & \text{if} \quad i \in S_j, \\ 0 & \text{if} \quad i \notin S_j. \end{cases}$$

It will also be useful to identify columns of X directly; hence we denote the ith column of X by x_i. Using X, we can easily express $\mathrm{cut}(S_1, S_2)$ as

$$\mathrm{cut}(S_1, S_2) = x_1^T A x_2 = \frac{1}{2}\langle X, AXB \rangle, \tag{1}$$

where

$$B = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

In [17] it is shown that an $n \times 3$ matrix X represents a partition of $V(G)$ into subsets S_1, S_2, and S_3 of prescribed sizes $m = (m_1, m_2, m_3)^T$ if and only if X satisfies the following relations:

$$X^T X = \mathrm{Diag}(m) =: M, \tag{2}$$

$$Xu_3 = u_n, \tag{3}$$

$$X \geq 0. \tag{4}$$

Note in particular that the constraint

$$X^T u_n = m, \tag{5}$$

asking that each partition block has the right number of elements, is implied by these conditions. The set of all $n \times 3$ matrices, representing some partition of $V(G)$ into sets of cardinalities, specified by m, will be denoted by $\mathcal{F}$. Using the above characterization of such partition matrices, we have

$$\mathcal{F} = \{X \in R^{n\times 3};\ X \text{ satisfies (2)–(4)}\}.$$

The MCP can equivalently be written as a quadratic program:

$$(MC_{QP}) \qquad \min\ \tfrac{1}{2}\langle X,\, AXB\rangle \text{ such that } X \in \mathcal{F}.$$

This problem has a nonconvex objective function, defined over a finite set. Our main goal in this section is to transform this problem into an equivalent linear program over the cone of completely positive matrices. We do this by expressing the linear constraints in an appropriate way as quadratic ones. Then we linearize the resulting quadratic terms. Specifically, we consider the following equations in the variable $X \in R^{n\times 3}$:

$$(e_i^T X u_3)^2 = \left(\sum_k X_{ik}\right)^2 = 1,\ 1 \leq i \leq n, \tag{6}$$

$$(u_n^T X e_j)(e_i^T X u_3) = \left(\sum_k X_{kj}\right)\left(\sum_k X_{ik}\right) = m_j,\ 1 \leq i \leq n,\ 1 \leq j \leq 3, \tag{7}$$

$$(u_n^T X e_i)(u_n^T X e_j) = \left(\sum_k X_{ki}\right)\left(\sum_k X_{kj}\right) = m_i m_j,\ 1 \leq i \leq j \leq 3. \tag{8}$$

Equations (6) are obtained by squaring the equations from (3). The equations (7) are obtained by elementwise multiplication of (3) and (5). The last set of equations is obtained from pairwise multiplication of (5). Clearly, any $X \in \mathcal{F}$ will satisfy (6)–(8). Using the Kronecker product and the property $\mathrm{vec}(PXQ) = (Q^T \otimes P)\,\mathrm{vec}(X)$ we get

$$\langle X,\, PXQ\rangle = \mathrm{vec}(X)^T\,\mathrm{vec}(PXQ) = x^T(Q^T \otimes P)x = \langle Q^T \otimes P,\, xx^T\rangle.$$

This helps us to reformulate the constraints (6)–(8) as follows:

$$\left\{\begin{array}{rcccl} (e_i^T X u_3)^2 & = & \langle X,\, e_i e_i^T X u_3 u_3^T\rangle & = & \langle J_3 \otimes E_{ii},\, xx^T\rangle, \\ (u_n^T X e_i)(e_j^T X u_3) & = & \langle X,\, u_n e_j^T X u_3 e_i^T\rangle & = & \langle V_i \otimes W_j^T,\, xx^T\rangle, \\ (u_n^T X e_i)(u_n^T X e_j) & = & \langle X,\, u_n u_n^T X e_j e_i^T\rangle & = & \langle E_{ij} \otimes J_n,\, xx^T\rangle. \end{array}\right. \tag{9}$$

In the last term we may replace E_{ij} with B_{ij}, since xx^T is symmetric. Similarly, we can rewrite the $(i,\, j)$th component on the left-hand side of (2) as

$$e_i^T X^T X e_j \quad = \quad \langle Xe_i,\, Xe_j\rangle \quad = \quad \langle X,\, XE_{ji}\rangle \quad = \quad \langle E_{ij} \otimes I,\, xx^T\rangle.$$

Let us now introduce $Y = xx^T$. Then MC_{QP} can be equivalently formulated as follows:

$$\begin{aligned} \min \quad & \frac{1}{2}\langle B^T \otimes A, Y\rangle \\ & \langle B_{ij} \otimes I, Y\rangle = m_i\delta_{ij}, \quad 1 \le i \le j \le 3, \qquad (10)\\ \text{s. t.} \quad & \langle J_3 \otimes E_{ii}, Y\rangle = 1, \quad 1 \le i \le n, \qquad (11)\\ & \langle V_i \otimes W_j^T, Y\rangle = m_i, \quad 1 \le i \le 3,\ 1 \le j \le n, \qquad (12)\\ & \langle B_{ij} \otimes J_n, Y\rangle = m_i m_j, \quad 1 \le i \le j \le 3, \qquad (13)\\ & Y = xx^T, \quad x \in R_+^{3n}. \end{aligned}$$

To see that this optimization problem is equivalent to MC_{QP}, we note that for any X feasible for MC_{QP}, we can take $x = \mathrm{vec}(X)$ to get a feasible $Y = xx^T$ for this problem with the same objective value and vice versa. The above problem has linear objective, linear constraints, and the quadratic equation, coupling Y and x. As a final simplification, we replace the constraints $Y = xx^T$ and $x \ge 0$ by $Y \in \mathcal{C}_{3n}^*$. The new optimization problem, which is a copositive program, will be denoted by MC_{CP}:

$$(MC_{CP}) \qquad \min \frac{1}{2}\langle B^T \otimes A, Y\rangle \text{ such that } Y \in \mathcal{C}_{3n}^* \text{ satisfies (10)–(13).}$$

The following theorem explains the relation between the feasible sets of MC_{QP} and MC_{CP}.

THEOREM 1.

$$\begin{aligned} &\mathrm{CONV}\{xx^T;\ x \in R_+^{3n},\ xx^T \textit{ feasible for } (10)\text{–}(13)\} \\ &= \{Y \in \mathcal{C}_{3n}^*;\ Y \textit{ feasible for } (10)\text{–}(13)\}. \end{aligned}$$

Proof. The "$\subseteq$" inclusion is obvious. To show inclusion in the other direction, we have to prove that for any $Y \in \mathcal{C}_{3n}^*$, feasible for MC_{CP}, there exist finitely many vectors $y^1, y^2, \ldots \in R_+^{3n}$ and numbers $\lambda_k \in [0, 1]$ with $\sum_k \lambda_k = 1$ such that $y^k(y^k)^T$ are feasible for constraints (10)–(13) and $Y = \sum_k \lambda_k y^k(y^k)^T$. Let $Y \in \mathcal{C}_{3n}^*$. From the definition of the cone $\mathcal{C}_{3n}^*$ follows that there exist finitely many nonzero vectors $x^k \in R_+^{3n}$ such that $Y = \sum_k x^k(x^k)^T$. We can treat x^k as a vector representation of some matrix $X^k \in R^{n\times 3}$; therefore we will index the components of each x^k with two indices: $x^k = (x_{ij}^k)$, $i = 1, \ldots, n$ and $j = 1, 2, 3$ (components $x_{:1}^k$ are the first n components of x^k—the first "column" of x^k, etc.). Let us first fix i and j ($1 \le i \le n$, $1 \le j \le 3$). If we denote with $r_k = \sum_{s=1}^3 x_{is}^k$ the sum of the "ith row" of x^k and with $c_k = \sum_{s=1}^n x_{sj}^k$ the sum of the "jth column" of x^k, then we can rewrite the constraints (11)–(13) using (6)–(9) as

$$\textstyle\sum_k r_k^2 \;=\; 1,\ \sum_k r_k c_k \;=\; m_j,\ \sum_k c_k^2 \;=\; m_j^2.$$

The Cauchy inequality, applied to vectors $v_1 = (r_1, r_2, \ldots)$ and $v_2 = (c_1, c_2, \ldots)$, implies $r_k = c_k/m_j$, or equivalently

$$\sum_s x_{is}^k = \frac{\sum_s x_{sj}^k}{m_j}, \quad k = 1, 2, \ldots. \qquad (14)$$

Since this is true for all i and j, we can see that the numbers $\sum_s x_{sj}^k/m_j$ are equal for all j. This means that in any vector x^k the sum of any "row" is equal to the sum

of column j divided by m_j for all $j = 1, 2, 3$. Therefore we may take without loss of generality $j = 1$ and define $\alpha_k = \sum_s x^k_{s1}/m_1$. Since none of x^k is zero we have $\alpha_k > 0$ for all k, and we may define $\lambda_k = \alpha_k^2 = (\sum_s x^k_{s1})^2/m_1^2$ and $y^k = x^k/\alpha_k$. From (13) we get

$$\sum_k \lambda_k = \frac{1}{m_1^2} \sum_k \Big(\sum_s x^k_{s1} \Big)^2 = 1.$$

Equation (14) implies that $y^k(y^k)^T$ are feasible for (11)–(13) and $Y = \sum_k \lambda_k y^k (y^k)^T$. To finish the proof it remains to show that $y^k(y^k)^T$ is feasible for (10) for all k. Indeed, if there exist $i \neq j$ and k such that $\langle B_{ij} \otimes I,\, y^k(y^k)^T \rangle > 0$, then because of nonnegativity of y^k we have $\langle B_{ij} \otimes I,\, Y \rangle > 0$, but this is a contradiction to the feasibility of Y. In particular, this means that in each "row" of y^k there is only one nonzero component, which must be equal to 1 because of feasibility for (11). Hence y^k is a 0–1 vector. This implies together with (13) that $\langle E_{ii} \otimes I,\, y^k(y^k)^T \rangle = \sum_s (y^k_{si})^2 = \sum_s y^k_{si} = m_i$; hence $y^k(y^k)^T$ is feasible for (10), too. □

The feasible set of MC_{CP} is therefore a polytope, spanned by the rank 1 matrices of type xx^T, where x is a vector representation of matrix X, feasible for MC_{QP}. Since MC_{CP} is a linear program, it has a rank 1 optimal solution; hence $OPT_{CP} \geq OPT_{QP}$. The opposite direction is obvious; hence we have the following corollary.

COROLLARY 2. *Problems MC_{QP} and MC_{CP} have the same optimal value; therefore the MCP can be equivalently formulated as a linear program in completely positive matrices.*

REMARK 2. *This copositive representation again confirms the importance of copositive programming in combinatorial optimization which was revealed by de Klerk and Pasechnik* [11]*, who proved that computing the stability number of a graph is equivalent to solving a copositive program and then presented a hierarchy of positive semidefinite relaxations, which follow from this approach and are strongly connected with the ϑ-function.*

3. The spectral relaxation as a semidefinite program. Helmberg et al. have derived in [10] a lower bound for OPT_{MC} which is easy to compute. They have omitted the nonnegativity constraint (4) in MC_{QP} and added constraint (5), yielding the problem

$$OPT_{HW} \quad = \quad \min \ \tfrac{1}{2}\langle X,\, \hat{A}XB \rangle \ \text{ such that } X \text{ satisfies (2), (3), and (5).}$$

In the above formulation we introduced $\hat{A} = A + D$ with $D = \frac{s(A)}{n} I - \mathrm{Diag}(r(A))$ and $s(A) = u^T A u$, $r(A) = Au$. This is a quadratic problem defined over a nonconvex set described by linear and quadratic equations. If we replace in the models MC_{QP} and MC_{CP} matrix A with $\hat{A}$, then the optimal values of these models do not change, since matrix XBX^T in the model MC_{QP} has only zeros on the main diagonal and similarly any feasible matrix Y in model MC_{CP} has only zeros on the main diagonals of off-diagonal blocks, as follows from (10) and complete positiveness of Y. Therefore $OPT_{MC} \geq OPT_{HW}$. Helmberg et al. [10] have in fact shown that OPT_{HW} has the explicit form

$$OPT_{HW} = -\frac{1}{2}\mu_2\lambda_2 - \frac{1}{2}\mu_1\lambda_n, \tag{15}$$

where λ_2 and λ_n are second smallest and the largest Laplacian eigenvalues of the graph G (i.e., the eigenvalues of matrix $L = \mathrm{Diag}(r(A)) - A = \frac{s(A)}{n} I - \hat{A}$) and $\mu_1 \geq \mu_2$ are defined as

$$\mu_{1,2} = \frac{1}{n}\Big(-m_1 m_2 \pm \sqrt{m_1 m_2 (n-m_1)(n-m_2)}\Big). \tag{16}$$

The key tool to get this result was the Hoffman–Wielandt inequality [9] combined with a projection technique for partitioning the nodes of a graph from [17]. It is an attractive feature of this bound that the closed form solution (15) is quite easy to compute, as it involves only the computation of the extreme Laplacian eigenvalues. On the other hand, the relaxation OPT_{HW}, as described above, does not permit the inclusion of further constraints, like, for instance, $X \geq 0$, without losing tractability. One of the main motivations for the current research was in fact the search for a new equivalent formulation of OPT_{HW} which is suitable for further tractable refinements. We now propose such a refinement. As already mentioned, we do not change the optimal value by replacing A with $\hat{A}$ in the models MC_{QP} and MC_{CP}. Let us consider the model, obtained from MC_{CP} by this replacement and relaxing the constraint $Y \in \mathcal{C}^*_{3n}$ to $Y \in \mathcal{S}^+_{3n}$. We will denote it by MC_{SDP} and its optimal value by OPT_{SDP}. Hence

$$\begin{array}{lrll} & OPT_{SDP} & = & \min \quad \frac{1}{2}\langle B^T \otimes \hat{A},\, Y\rangle \\ (MC_{SDP}) & & & \text{s. t.} \quad Y \in \mathcal{S}^+_{3n} \\ & & & \qquad\quad Y \text{ satisfies (10)–(13).} \end{array}$$

In the next section we will show that the value OPT_{SDP} is equal to OPT_{HW}. First we have the easy part.

LEMMA 3.

$$OPT_{HW} \geq OPT_{SDP}.$$

Proof. If X satisfies (2), (3), and (5), then X satisfies constraints (2) and (6)–(8). The matrix $Y = xx^T$ satisfies (10)–(13) and is in $\mathcal{S}^+_{3n}$ hence is feasible for MC_{SDP}. Since $\frac{1}{2}\langle X,\, \hat{A}XB\rangle = \frac{1}{2}\langle B^T \otimes \hat{A},\, Y\rangle$, the lemma follows. ∎

The main result of this section is the following theorem.

THEOREM 4.

$$OPT_{HW} = OPT_{SDP}.$$

From Lemma 3 it follows that we need to prove that $OPT_{HW} \leq OPT_{SDP}$. Our proof of this result is rather involved and consists of two major steps. In the first step we reformulate the semidefinite program MC_{SDP} in a new coordinate system, obtained by diagonalizing the cost matrix $B^T \otimes \hat{A}$. This is the content of subsection 4.1, which ends with the main result of the first step, i.e., with the semidefinite program (18). The second part of the proof is more subtle. We extract a subproblem of (18) by leaving out some constraints and projecting the feasible set to a proper hyperplane. In subsection 4.2 we show that this subproblem, denoted by MC_{SDPa}, in fact captures the essential part of MC_{SDP}, and its optimal solution OPT_{SDPa} satisfies $OPT_{SDP} \geq OPT_{SDPa} \geq OPT_{HW}$. This closes the chain of inequalities and the proof is finished.

4. Proof of Theorem 4.

4.1. Diagonalization of the cost matrix. Let $\frac{1}{2}\hat{A} = PSP^T$, $B = QTQ^T$, where P and Q are orthonormal matrices whose columns are eigenvectors of $\frac{1}{2}\hat{A}$ and B, respectively, and S, T are diagonal matrices with eigenvalues on the diagonal. We

take the factorizations where the eigenvalues are in nondecreasing order; hence we have

$$Q = \frac{1}{2}\begin{bmatrix} -\sqrt{2} & 0 & \sqrt{2} \\ \sqrt{2} & 0 & \sqrt{2} \\ 0 & 2 & 0 \end{bmatrix}, \; T = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

If we denote with $\ell_i = s_{ii}$, then from $\hat{A} = \frac{s(A)}{n} I - L$ (see the beginning of the previous section) follows $\ell_i = \frac{s(A)}{2n} - \frac{\lambda_{n-i+1}(L)}{2}$, in particular $\ell_1 = \frac{s(A)}{2n} - \frac{\lambda_n(L)}{2}$ and $\ell_n = \frac{s(A)}{2n}$. We choose P in such a way that the last column of P is equal to $u/\sqrt{n}$. This can be done since u is an eigenvector of $\hat{A}$ corresponding to the largest eigenvalue of $\hat{A}$. In the following lemmas we investigate what happens if we substitute in the model MC_{SDP} the matrix variable Y with matrix variable Z, which are related by

$$Y = (Q \otimes P)\, Z\, (Q \otimes P)^T. \tag{17}$$

This substitution simplifies the objective function, which becomes $\langle T \otimes S,\, Z\rangle$; hence only diagonal elements of Z will determine the objective value. If $Y \in \mathcal{S}^+_{3n}$, then the new matrix variable Z is from $\mathcal{S}^+_{3n}$, too. We will often for the sake of simplicity write matrix Z as a block matrix: $Z = [Z^{ij}]_{1\le i,j\le 3}$, where $Z^{ij} \in R^{n\times n}$. This actually means that

$$Z = \sum_{1\le i,j\le 3} E_{ij} \otimes Z^{ij} = \begin{bmatrix} Z^{11} & Z^{12} & Z^{13} \\ Z^{21} & Z^{22} & Z^{23} \\ Z^{31} & Z^{32} & Z^{33} \end{bmatrix}.$$

We will denote with Z^{ij}_{kl} the (k,l)th component of matrix Z^{ij}.

LEMMA 5. *Let $Y, Z \in \mathcal{S}_{3n}$ satisfy* (17). *The matrix Y satisfies constraint* (13) *if and only if the matrix Z satisfies*

$$Z^{ij}_{nn} = f_{ij},\; 1 \le i \le j \le 3, \tag{13a}$$

where matrix $F = (f_{ij}) \in \mathcal{S}^+_3$ is as follows:

$$F = \frac{1}{2n}\begin{bmatrix} m_2 - m_1 \\ \sqrt{2}m_3 \\ m_2 + m_1 \end{bmatrix} \cdot \begin{bmatrix} m_2 - m_1 \\ \sqrt{2}m_3 \\ m_2 + m_1 \end{bmatrix}^T.$$

Proof. Here we use the fact that $P(:,n) = u/\sqrt{n}$. Constraint (13) becomes $\langle (Q^T B_{ij} Q) \otimes (P^T J_n P),\, Z\rangle = m_i m_j$. Since all columns of P are orthogonal, we have $P^T J_n P = P^T W_n^T = nE_{nn}$. We also get matrices $\tilde{B}_{ij} := Q^T B_{ij} Q$:

$$\begin{aligned}
\tilde{B}_{11} &= \tfrac{1}{2}\begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix}, & \tilde{B}_{12} &= \tfrac{1}{2}\begin{bmatrix} -1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \\
\tilde{B}_{13} &= \tfrac{\sqrt{2}}{4}\begin{bmatrix} 0 & -1 & 0 \\ -1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, & \tilde{B}_{22} &= \tfrac{1}{2}\begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{bmatrix}, \\
\tilde{B}_{23} &= \tfrac{\sqrt{2}}{4}\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, & \tilde{B}_{33} &= \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}.
\end{aligned}$$

Therefore we get $\langle B_{11} \otimes J_n,\, Y\rangle = n\langle \tilde{B}_{11} \otimes E_{nn},\, Z\rangle = \frac{n}{2}(Z^{11}_{nn} - 2Z^{13}_{nn} + Z^{33}_{nn})$. Equation $\langle B_{11} \otimes J_n,\, Y\rangle = m_1^2$ is thus equivalent to

$$Z^{11}_{nn} - 2Z^{13}_{nn} + Z^{33}_{nn} = \frac{2m_1^2}{n}.$$

Similarly, we rewrite the other equations from constraint (13) into

$$\begin{aligned} -Z^{11}_{nn} + Z^{33}_{nn} &= \tfrac{2m_1m_2}{n}, & -Z^{12}_{nn} + Z^{23}_{nn} &= \tfrac{\sqrt{2}m_1m_3}{n},\\ Z^{11}_{nn} + 2Z^{13}_{nn} + Z^{33}_{nn} &= \tfrac{2m_2^2}{n}, & Z^{12}_{nn} + Z^{23}_{nn} &= \tfrac{\sqrt{2}m_2m_3}{n},\\ Z^{22}_{nn} &= \tfrac{m_3^2}{n}. \end{aligned}$$

The solution of this system of six linear equations in six variables is $Z^{ij}_{nn} = f_{ij}$. ∎

LEMMA 6. *Let $Y, Z \in \mathcal{S}_{3n}$ satisfy* (17). *The matrix Y satisfies constraint* (10) *if and only if the matrix Z satisfies constraint*

$$\operatorname{trace}(Z^{ij}) = h_{ij},\ 1 \le i \le j \le 3, \tag{10a}$$

where matrix $H = (h_{ij}) \in \mathcal{S}_3$ is defined as

$$H = \frac{1}{2}\begin{bmatrix} m_1 + m_2 & 0 & m_2 - m_1 \\ 0 & 2m_3 & 0 \\ m_2 - m_1 & 0 & m_1 + m_2 \end{bmatrix}.$$

Proof. From $P^T I\, P = I$ follows $\langle B_{ij} \otimes I,\, Y\rangle = \langle \tilde{B}_{ij} \otimes I,\, Z\rangle$. If $i = j = 1$, then $\langle \tilde{B}_{11} \otimes I,\, Z\rangle = (\operatorname{trace}(Z^{11}) - 2\operatorname{trace}(Z^{13}) + \operatorname{trace}(Z^{33}))/2$, so the first equation from (10) could be rewritten as

$$\frac{1}{2}\Big(\operatorname{trace}(Z^{11}) - 2\operatorname{trace}(Z^{13}) + \operatorname{trace}(Z^{33})\Big) = m_1.$$

Similarly, we get the other five linear equations in six variables $\operatorname{trace}(Z^{ij})$. The unique solution is given by $\operatorname{trace}(Z^{ij}) = h_{ij}$, $1 \le i \le j \le 3$. ∎

LEMMA 7. *Let $Y, Z \in \mathcal{S}_{3n}$ satisfy* (17).

(a) *The matrix Y satisfies constraint* (11) *if and only if the matrix Z satisfies the constraint*

$$\big\langle U \otimes P(i,:)^T P(i,:),\, Z\big\rangle = 1,\ 1 \le i \le n, \tag{11a}$$

where

$$U = Q^T J_3 Q = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & \sqrt{2} \\ 0 & \sqrt{2} & 2 \end{bmatrix}.$$

(b) *The matrix Y satisfies constraint* (12) *if and only if the matrix Z satisfies the constraint*

$$\langle \tilde{V}_i \otimes (e_n \cdot P(j,:)),\, Z\rangle = \frac{m_i}{\sqrt{n}},\ 1 \le i \le 3,\ 1 \le j \le n, \tag{12a}$$

where $\tilde{V}_i = Q^T V_i Q$.

Proof. (a) This statement follows immediately from $P^T E_{ii} P = P(i,:)^T P(i,:)$.

(b) After the substitution the left-hand side of constraint (12) becomes $\langle (Q^T V_i Q) \otimes (P^T W_j^T P),\, Z\rangle = m_i$. A short calculation shows

$$\tilde{V}_1 \;=\; \tfrac{1}{2}\begin{bmatrix} 0 & -\sqrt{2} & -2 \\ 0 & 0 & 0 \\ 0 & \sqrt{2} & 2 \end{bmatrix}, \quad \tilde{V}_2 \;=\; \tfrac{1}{2}\begin{bmatrix} 0 & \sqrt{2} & 2 \\ 0 & 0 & 0 \\ 0 & \sqrt{2} & 2 \end{bmatrix}, \quad \tilde{V}_3 \;=\; \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & \sqrt{2} \\ 0 & 0 & 0 \end{bmatrix}.$$

The term $P^T W_j^T P$ simplifies because of the choice of the last column of P into $\sqrt{n} E_{nj} P = \sqrt{n} e_n e_j^T P = \sqrt{n} e_n P(j,:)$. ∎

By introducing the set

$$\mathcal{G} = \{Z \in \mathcal{S}^+_{3n},\ Z \text{ satisfies constraints (10a), (11a), (12a), and (13a)}\},$$

we can see that the problem MC_{SDP} is equivalent to the problem

$$\min \langle T \otimes S,\, Z\rangle \ \text{ such that } Z \in \mathcal{G}, \tag{18}$$

since for any feasible solution Y for MC_{SDP} we can find a solution $Z \in \mathcal{G}$ via (17) with the same value of the objective value and vice versa. It should be noted that the cost function in (18) simplifies to $\langle T \otimes S,\, Z\rangle = \sum_{i=1}^n \ell_i (Z_{ii}^{33} - Z_{ii}^{11})$.

4.2. A block-diagonal subproblem. The semidefinite program (18) is still quite complicated. Since Lemmas 5–7 show that feasibility for constraints (10a)–(13a) is mostly determined with the diagonal entries of blocks Z^{ij}, we are going to study the following semidefinite program, which we obtain by keeping in the program (18) only constraints (10a) and (13a) and ignoring all nondiagonal components in any block Z^{ij}. We also omit blocks Z^{2i} and Z^{i2}, $i = 1, 2, 3$, since they do not contribute to the cost function.

$$(MC_{SDPa}) \qquad \begin{array}{lrcl} \min & \sum_{i=1}^{n-1} \ell_i (r_i - p_i) + \ell_n (f_{33} - f_{11}) & & \\ \text{s. t.} & \sum_{i=1}^{n-1} p_i & = & h_{11} - f_{11} := b_1, \\ & \sum_{i=1}^{n-1} r_i & = & h_{33} - f_{33} := b_2, \\ & \sum_{i=1}^{n-1} q_i & = & h_{13} - f_{13} := b_3, \\ & U_i & = & \begin{bmatrix} p_i & q_i \\ q_i & r_i \end{bmatrix} \succeq 0. \end{array}$$

The constants b_i are

$$b_1 \;=\; \frac{4m_1m_2 + m_1m_3 + m_2m_3}{2n}, \quad b_2 \;=\; \frac{(m_1+m_2)m_3}{2n}, \quad \text{and} \quad b_3 \;=\; \frac{(m_2-m_1)m_3}{2n}.$$

In the following lemma we compare the optimal values of MC_{SDP} and MC_{SDPa}.

LEMMA 8.

$$OPT_{SDP} \geq OPT_{SDPa}.$$

Proof. We will show that any feasible solution for (18) implies a feasible solution for MC_{SDPa}. Let $Z = [Z^{ij}]$ be a feasible solution for (18) and let us define $p_i = Z_{ii}^{11}$, $r_i = Z_{ii}^{33}$, and $q_i = Z_{ii}^{13}$ for $1 \leq i \leq n-1$. From Lemmas 5 and 6 follows $\sum_{i=1}^{n-1} p_i = \text{trace}(Z^{11}) - Z_{nn}^{11} = b_1$, and hence p_i are feasible for the first equation in MC_{SDPa}. Similarly we can show that the other two constraints are satisfied and

that the matrices $U_i = \left[\begin{smallmatrix} p_i & q_i \\ q_i & r_i \end{smallmatrix}\right]$ are positive semidefinite, following from $Z \succeq 0$. The objective value of the MC_{SDPa} is exactly $\sum_{i=1}^{n} \ell_i(Z_{ii}^{11} - Z_{ii}^{33}) = \langle T \otimes S, Z\rangle$; hence the lemma follows. □

Here is the dual semidefinite program for MC_{SDPa}:

$$\begin{aligned} & \max \quad b_1y_1 + b_2y_2 + 2b_3y_3 + \ell_n(f_{33} - f_{11}) \\ (DMC_{SDPa}) \qquad & \text{s. t.} \quad V_i = \begin{bmatrix} -\ell_i - y_1 & -y_3 \\ -y_3 & \ell_i - y_2 \end{bmatrix} \succeq 0, \quad 1 \le i \le n-1. \end{aligned}$$

First let us introduce the number

$$\delta = \frac{2m_1m_2 + m_1m_3 + m_2m_3}{2\sqrt{m_1m_2(n-m_1)(n-m_2)}} = \frac{m_1(n-m_1) + m_2(n-m_2)}{2\sqrt{m_1m_2(n-m_1)(n-m_2)}}.$$

This number is well defined in view of Remark 1. Note also that δ is of the form

$$\frac{1}{2}\left(u + \frac{1}{u}\right) \text{ with } u = \sqrt{\frac{m_1(n-m_1)}{m_2(n-m_2)}} > 0.$$

Therefore $\delta \ge 1$. The next lemma allows us to finish the proof of Theorem 4. We need the following simple observation for its proof.

PROPOSITION 9. *If $a + b = c + d$ and $|a - b| \le |c - d|$, then $ab \ge cd$.*

Proof. We can write $(a-b)^2 = (a+b)^2 - 4ab$ and $(c-d)^2 = (c+d)^2 - 4cd$. Using the assumptions of the proposition we get $(a+b)^2 - 4ab \ge (c+d)^2 - 4cd$ and the result follows. □

LEMMA 10. *The numbers*

$$\begin{aligned} y_1 &= -\frac{\ell_1 + \ell_{n-1}}{2} - \frac{\delta}{2}(\ell_{n-1} - \ell_1), \\ y_2 &= \frac{\ell_1 + \ell_{n-1}}{2} - \frac{\delta}{2}(\ell_{n-1} - \ell_1), \\ y_3 &= \sqrt{(-\ell_1 - y_1)(\ell_1 - y_2)} \end{aligned}$$

form an optimal solution for the dual problem DMC_{SDPa} with objective value equal to OPT_{HW}.

Proof. First note that $\delta \ge 1$ implies $y_1 \le -\ell_{n-1}$ and $y_2 \le \ell_1$. This shows that in the definition of y_3 we take the square root of a nonnegative number; hence y_3 is well defined. To see that $V_i \succeq 0$, we first note that the numbers ℓ_i are in nondecreasing order; therefore $-\ell_i - y_1 \ge 0$, $\ell_i - y_2 \ge 0$. Using that $y_2 = y_1 + \ell_1 + \ell_{n-1}$ we get $y_3^2 = (-\ell_1 - y_1)(\ell_1 - y_2) = (-\ell_{n-1} - y_1)(\ell_{n-1} - y_2)$ and $(-\ell_i - y_1) + (\ell_i - y_2) = -y_1 - y_2 = \delta(\ell_{n-1} - \ell_1)$. Since $|(-\ell_i - y_1) - (\ell_i - y_2)| = |\ell_1 + \ell_{n-1} - 2\ell_i| \le |\ell_{n-1} - \ell_1| = |(-\ell_1 - y_1) - (\ell_1 - y_2)|$, we get by Proposition 9

$$(-\ell_i - y_1)(\ell_i - y_2) \ge (-\ell_1 - y_1)(\ell_1 - y_2) = (-\ell_{n-1} - y_1)(\ell_{n-1} - y_2) = y_3^2;$$

hence $\det(V_i) \ge 0$ and positive semidefiniteness of V_i follows. Second we will show the optimality of (y_1, y_2, y_3). It is sufficient to prove that

$$b_1y_1 + b_2y_2 + 2b_3y_3 + \ell_n(f_{33} - f_{11}) = -\frac{1}{2}\mu_2\lambda_2 - \frac{1}{2}\mu_1\lambda_n = OPT_{HW},$$

since from Lemmas 3 and 8 and the weak duality property we know that the optimal value of DMC_{SDPa} is at most $OPT_{SDP} \le OPT_{HW}$. Using the fact that

$\ell_1 = \frac{s(A)}{2n} - \frac{\lambda_n}{2}$, $\ell_{n-1} = \frac{s(A)}{2n} - \frac{\lambda_2}{2}$, and $\ell_n(f_{33} - f_{11}) = \frac{s(A)m_1m_2}{n^2}$, it remains to show that

$$b_1y_1 + b_2y_2 + 2b_3y_3 = \mu_2\ell_{n-1} + \mu_1\ell_1.$$

One can derive that

$$y_3 = \sqrt{(-\ell_1 - y_1)(\ell_1 - y_2)} = \frac{\ell_{n-1}-\ell_1}{2}\sqrt{\delta^2 - 1} = \frac{m_3(m_2-m_1)(\ell_{n-1}-\ell_1)}{4\sqrt{m_1m_2(n-m_1)(n-m_2)}}$$

and show

$$\begin{aligned} b_1y_1 + b_2y_2 + 2b_3y_3 &= \frac{\ell_1}{2}(\delta(b_1+b_2) - b_1 + b_2 - 2b_3\frac{m_3(m_2-m_1)}{2\sqrt{m_1m_2(n-m_1)(n-m_2)}}) \\ &\quad - \frac{\ell_{n-1}}{2}(\delta(b_1+b_2) + b_1 - b_2 - 2b_3\frac{m_3(m_2-m_1)}{2\sqrt{m_1m_2(n-m_1)(n-m_2)}}) \\ &= \mu_1\ell_1 + \mu_2\ell_{n-1}. \end{aligned}$$

Checking the last equality involves tedious but straightforward algebraic manipulation. □

Proof of Theorem 4. From Lemmas 3, 8, and 10 and the weak duality property for semidefinite program MC_{SDPa} follows

$$OPT_{HW} \geq OPT_{SDP} \geq OPT_{SDPa} \geq OPT_{DSDPa} = OPT_{HW};$$

hence equality holds throughout. □

4.3. Reconstructing the optimal solution of the problem MC_{SDP}. Once we know the optimal solution of the dual problem DMC_{SDPa}, we can reconstruct the optimal solution of MC_{SDP} by tracing the procedure from the previous subsection and using the structural information about the feasible set $\mathcal{G}$. We will first compute the optimal solution of MC_{SDPa} from the optimal solution of MC_{DSDPa} and then will extend it to the optimal solution of MC_{SDP}. Let $U^* = \mathrm{diag}(U_1, \ldots, U_{n-1})$ be the optimal solution of MC_{SDPa} and (y_1, y_2, y_3) the optimal solution for DMC_{SDPa} from Lemma 8. We define matrix $V^* = \mathrm{diag}(V_1, \ldots, V_{n-1})$ with

$$V_i = \begin{bmatrix} -\ell_i - y_1 & -y_3 \\ -y_3 & \ell_i - y_2 \end{bmatrix}, \quad 1 \leq i \leq n-1. \tag{19}$$

From the feasibility of (y_1, y_2, y_3) it follows that $V_i \succeq 0$ and any matrix V_i is in fact the dual matrix to U_i for $1 \leq i \leq n-1$. Since V^* is actually optimal for DMC_{SDPa}, the strong duality property implies $\langle U_i, V_i \rangle = 0$ for $1 \leq i \leq n-1$. Suppose first that $\ell_1 < \ell_{n-1}$ and V_1 and V_{n-1} are the only singular matrices in V^* (hence V_1 and V_{n-1} are rank one matrices). Let

$$U_1 = \begin{bmatrix} p_1 & q_1 \\ q_1 & r_1 \end{bmatrix}, \; U_{n-1} = \begin{bmatrix} p_2 & q_2 \\ q_2 & r_2 \end{bmatrix}, \; V_1 = \begin{bmatrix} v_1 & z_1 \\ z_1 & w_1 \end{bmatrix}, \text{ and } V_{n-1} = \begin{bmatrix} v_2 & z_2 \\ z_2 & w_2 \end{bmatrix}.$$

Using (19) we see that $v_1 = -\ell_1 - y_1$, $z_1 = -y_3$, $w_1 = \ell_1 - y_2$, etc. From the strong duality property it follows that $U_2, U_3, \ldots, U_{n-2}$ are zero matrices and U_1, U_{n-1} are singular. Since U_1, U_{n-1}, V_1, and V_{n-1} are singular, the following must be true:

$$\begin{aligned} z_1^2 &= v_1w_1, \quad z_2^2 = v_2w_2, \\ q_1^2 &= p_1r_1, \quad q_2^2 = p_2r_2. \end{aligned}$$

Together with the strong duality property $\langle U_1, V_1\rangle = p_1v_1 + 2q_1z_1 + r_1w_1 = 0$ this implies that

$$\frac{p_1v_1 + r_1w_1}{2} = |q_1z_1| = \sqrt{p_1v_1r_1w_1}.$$

From the arithmetic-geometric inequality it follows that $p_1v_1 = r_1w_1$ and similarly $p_2v_2 = r_2w_2$. Components of U_1 and U_{n-1} must also satisfy linear constraints from MC_{SDPa}: $p_1 + p_2 = b_1$, $r_1 + r_2 = b_2$, and $q_1 + q_2 = b_3$. All these equations uniquely determine the components of U_1 and U_{n-1} as

$$\begin{array}{cccccccccc} p_1 & = & \alpha w_1, & q_1 & = & -\alpha z_1, & r_1 & = & \alpha v_1, \\ p_2 & = & \beta w_2, & q_2 & = & -\beta z_2, & r_2 & = & \beta v_2, \end{array} \tag{20}$$

where

$$\begin{array}{rcccc} \alpha & = & \frac{b_2w_2 - b_1v_2}{v_1w_2 - v_2w_1} = \frac{-m_1m_2 + \sqrt{m_1m_2(n-m_1)(n-m_2)}}{(\ell_{n-1} - \ell_1)n} & = & \frac{\mu_1}{\ell_{n-1} - \ell_1}, \\ \beta & = & \frac{b_1v_1 - b_2w_1}{v_1w_2 - v_2w_1} = \frac{m_1m_2 + \sqrt{m_1m_2(n-m_1)(n-m_2)}}{(\ell_{n-1} - \ell_1)n} & = & -\frac{\mu_2}{\ell_{n-1} - \ell_1}. \end{array}$$

If we have $\ell_1 < \ell_{n-1}$ and there exists $1 < i < n-1$ such that V_i is a rank one matrix, then the matrix $U^* = \mathrm{diag}(U_1, \ldots, U_{n-1})$, where $U_2, \ldots, U_{n-2}$ are zero matrices and components of U_1 and U_{n-1} are those from (20), is still the (nonunique) optimal solution of MC_{SDPa}. The last case is that $\ell_1 = \ell_{n-1}$. In this case we cannot use U_1 and U_{n-1}, defined with (20), because α and β are not defined. We will find the optimal solution of MC_{SDPa} directly. Let us define U_1 and U_{n-1} with

$$p_1 = p_2 = \frac{b_1}{2}, \quad r_1 = r_2 = \frac{b_2}{2}, \quad q_1 = q_2 = \frac{b_3}{2}, \tag{21}$$

and let U_i be zero matrices for $2 \le i \le n-2$. The matrix $U = \mathrm{Diag}(U_1, \ldots, U_{n-1})$ is feasible for MC_{SDPa} and $\sum_{i=1}^{n-1} \ell_i(r_i - p_i) + \ell_n(f_{33} - f_{11}) = \ell_1(b_2 - b_1) + \ell_n(f_{33} - f_{11}) = -\frac{2m_1m_2\ell_1}{n} + \frac{s(A)m_1m_2}{n^2} = OPT_{HW}$; hence U is optimal for MC_{SDPa}. However, the MCP is trivial if $\ell_1 = \ell_{n-1}$, since in this case the underlying graph is the complete graph K_n. Let us introduce the matrices

$$Z_1 = \begin{bmatrix} p_1 & -\sqrt{2}q_1 & q_1 \\ -\sqrt{2}q_1 & 2r_1 & -\sqrt{2}r_1 \\ q_1 & -\sqrt{2}r_1 & r_1 \end{bmatrix}, \; Z_{n-1} = \begin{bmatrix} p_2 & -\sqrt{2}q_2 & q_2 \\ -\sqrt{2}q_2 & 2r_2 & -\sqrt{2}r_2 \\ q_2 & -\sqrt{2}r_2 & r_2 \end{bmatrix},$$

and $Z_n = F$, where $F \in \mathcal{S}_3^+$ is from Lemma 5, and p_i, r_i, and q_i are either from (20) or from (21).

PROPOSITION 11. *The matrix*

$$Z^* = Z_1 \otimes E_{11} + Z_{n-1} \otimes E_{n-1,n-1} + Z_n \otimes E_{nn} \tag{22}$$

is the optimal solution for (18) *and the matrix*

$$Y^* = (Q \otimes P)\, Z^*\, (Q \otimes P)^T$$

is the optimal solution for MC_{SDP}.

Proof. The structure of Z^* for the case $n = 3$ can be seen in Figure 1. From the construction of Z^*, Theorem 4, and Proposition 10 follows that $\langle T \otimes S, Z^*\rangle = \ell_1(r_1 - p_1) + \ell_{n-1}(r_2 - p_2) + \ell_n(f_{33} - f_{11}) = OPT_{HW} = OPT_{SDP}$; hence Z^* gives the optimal value of (18). Therefore it remains to show that Z^* is feasible for the problem (18). Positive semidefiniteness of Z^* follows from positive semidefiniteness

$$Z^* = \left[\begin{array}{ccc|ccc|ccc}
p_1 & 0 & 0 & -\sqrt{2}q_1 & 0 & 0 & q_1 & 0 & 0 \\
0 & p_2 & 0 & 0 & -\sqrt{2}q_2 & 0 & 0 & q_2 & 0 \\
0 & 0 & f_{11} & 0 & 0 & f_{12} & 0 & 0 & f_{13} \\
\hline
-\sqrt{2}q_1 & 0 & 0 & 2r_1 & 0 & 0 & -\sqrt{2}r_1 & 0 & 0 \\
0 & -\sqrt{2}q_2 & 0 & 0 & 2r_2 & 0 & 0 & -\sqrt{2}r_2 & 0 \\
0 & 0 & f_{12} & 0 & 0 & f_{22} & 0 & 0 & f_{23} \\
\hline
q_1 & 0 & 0 & -\sqrt{2}r_1 & 0 & 0 & r_1 & 0 & 0 \\
0 & q_2 & 0 & 0 & -\sqrt{2}r_2 & 0 & 0 & r_2 & 0 \\
0 & 0 & f_{13} & 0 & 0 & f_{23} & 0 & 0 & f_{33}
\end{array}\right]$$

FIG. 1. *Structure of Z^* for $n = 3$.*

of matrices U_1, U_{n-1}, and F. Feasibility for the constraints (10a) and (13a) follows immediately from the feasibility of U_1 and U_{n-1} for the problem MC_{SDPa} and the structure of Z^*.

To check the feasibility for (11a) we need to compute for all $1 \le i \le n$

$$\begin{aligned}
\left\langle U \otimes P(i,:)^T P(i,:),\, Z^* \right\rangle &= P(i,n)^2(f_{22} + 2\sqrt{2}f_{23} + 2f_{33}) \\
&\quad + (P(i,1)^2 + P(i,n-1)^2)(2r_1 + 2r_2 - 4r_1 - 4r_2 + 2r_1 + 2r_2) \\
&= (f_{22} + 2\sqrt{2}f_{23} + 2f_{33})/n \;+\; 0 \;=\; 1,
\end{aligned}$$

so Z^* is feasible for (11a). The last constraint (12a) reduces for $i = 1$ and arbitrary $1 \le j \le n$ to

$$\begin{aligned}
\langle \tilde{V}_1 \otimes e_n P(j,:),\, Z^* \rangle &= \frac{P(j,n)}{2}\left(-\sqrt{2}f_{12} - 2f_{13} + \sqrt{2}f_{23} + 2f_{33}\right) \\
&= \frac{1}{2\sqrt{n}} 2m_1 \;=\; \frac{m_1}{\sqrt{n}}.
\end{aligned}$$

Similarly, we check the feasibility for (12a) for $i = 2, 3$. Once we know that Z^* is optimal for (18), the optimality of Y^* follows from Lemmas 5–7 and the fact that $\langle T \otimes S,\, Z^* \rangle = \frac{1}{2}\langle B \otimes \hat{A},\, Y^* \rangle$. □

A simple implication of Proposition 11 is the following closed form formula for the optimal solution of the semidefinite program MC_{SDP}:

$$\begin{aligned}
Y^* = (Q \otimes P)\, Z^*\, (Q \otimes P)^T &= (QZ_1Q^T) \otimes (P(:,1)P(:,1)^T) \\
&\quad + (QZ_{n-1}Q^T) \otimes (P(:,n-1)P(:,n-1)^T) + \frac{1}{n}(QZ_nQ^T) \otimes J_n.
\end{aligned} \tag{23}$$

We can see that for any graph and fixed m, Y^* is completely determined by (y_1, y_2, y_3) from Lemma 8 and hence with the first and the second to last eigenvalues of $\hat{A}$ and corresponding eigenvectors, which are determined by the second and the last eigenvalues of the graph Laplacian (λ_2 and λ_n) and corresponding eigenvectors.

5. A new family of relaxations for the MCP. In the previous section we have seen that relaxing the constraint $Y \in \mathcal{C}^*_{3n}$ in model MC_{CP} to $Y \in \mathcal{S}^+_{3n}$ leads to the lower bound OPT_{HW}. To get a better lower bound it is therefore natural to use a (tractable) set $\mathcal{K}$ with $\mathcal{C}^*_{3n} \subset \mathcal{K} \subset \mathcal{S}^+_{3n}$. Specifically, let $OPT_{\mathcal{K}}$ be defined by

$$OPT_{\mathcal{K}} = \min \frac{1}{2}\langle B^T \otimes \hat{A},\, Y \rangle \text{ such that } Y \in \mathcal{K} \text{ and } Y \text{ satisfies (10)–(13)};$$

then $OPT_{MC} \ge OPT_{\mathcal{K}} \ge OPT_{HW}$. A simple (and tractable) candidate for the set $\mathcal{K}$ is $\mathcal{K}_0 = \mathcal{S}^+_{3n} \cap \mathcal{N}_{3n}$. This is actually the first member in the hierarchy of cones introduced by Parrilo in [15] and used also by de Klerk and Pasechnik in their work

about the stability number in [11]. We may also replace it with any other member of this hierarchy, but already the second cone $\mathcal{K}_1$ leads to a very expensive semidefinite program. $OPT_{\mathcal{K}_0}$ is already quite expensive, since each sign constraint contributes one linear equation and one slack variable and we have approximately $9n^2/2$ of them. We get cheaper models if we take for $\mathcal{K}$ the cone

$$\mathcal{K}_0^a = \{X \in \mathcal{S}_{3n}^+,\ \mathcal{Z}(X) = 0, \text{ and } X_{ij}^{12} \geq 0 \text{ for any } (i,j) \text{ with } a_{ij} > 0\},$$

where $\mathcal{Z}(X) = 0$ means that all diagonal entries in all nondiagonal blocks must be zero, which corresponds to componentwise orthogonality of columns of partition matrices. Taking the last cone makes sense, since the matrix $B^T \otimes \hat{A}$ in the model MC_{CP} is nonzero only in those positions of $(1,2)$th and $(2,1)$th blocks, where $a_{ij} > 0$, and the constraint $\mathcal{Z}(X) = 0$ is satisfied by any feasible solution for MC_{CP}. Table 1 shows numerical results, which we obtained by optimizing over the cones $\mathcal{K}_0$ and $\mathcal{K}_0^a$. Table 1 contains computational results on small graphs: $P_6 \times P_4$ is the product of two paths, i.e., a 6×4 grid graph, $K_{6,9}$ is the complete bipartite graph on 15 nodes, and rand(15, 0.5) is a random graph on 15 nodes with edge density 0.5. We partition them in several different ways, given by m in column 2. The vectors m are exactly those for which $m_2/2 \leq m_1 \leq m_2$ and m_3 fixed (later we will see that this is useful when considering the balanced vertex separators of a graph). For all these graphs except the random graph we can determine OPT_{MC} by inspection; see column 3. The last three columns contain the original bound OPT_{HW} from [10] and improvements obtained by optimizing over the cones $\mathcal{K}_0$ and $\mathcal{K}_0^a$.

TABLE 1
MCP and the relaxations on some small graphs.

Graph	m_1, m_2, m_3	OPT_{MC}	OPT_{HW}	$OPT_{\mathcal{K}_0}$	$OPT_{\mathcal{K}_0^a}$
$P_6 \times P_4$	7 14 3	1	-3.36	0.26	0.00
$P_6 \times P_4$	8 13 3	1	-3.32	0.22	0.00
$P_6 \times P_4$	9 12 3	1	-3.31	0.15	0.00
$P_6 \times P_4$	10 11 3	1	-3.29	0.10	0.00
$P_6 \times P_4$	8 14 2	2	-1.88	1.12	0.00
$P_6 \times P_4$	9 13 2	2	-1.84	1.09	0.00
$P_6 \times P_4$	10 12 2	2	-1.81	0.94	0.05
$P_6 \times P_4$	11 11 2	2	-1.80	0.85	0.06
rand(15, 0.5)	5 6 4	7	0.19	6.65	6.07
$K_{6,9}$	4 6 5	4	2.18	4.00	3.61
$K_{6,9}$	5 5 5	5	2.50	4.79	3.99
$K_{6,9}$	4 7 4	8	4.71	8.00	6.87
$K_{6,9}$	5 6 4	9	5.41	8.99	7.75

While the relaxation over $\mathcal{K}_0$ provides a substantial improvement as compared to OPT_{HW}, this bound is also rather expensive: we need to solve a semidefinite program in matrices of order $3n$ with approximately $9n^2/2$ additional constraints. Looking at $OPT_{\mathcal{K}_0^a}$ we see that this relaxation is slightly weaker than $OPT_{\mathcal{K}_0}$ but is less expensive since it includes only approximately $m = |E|$ additional constraints, if E is the edge set of the graph. When it is positive, then it is significantly better than OPT_{HW}. We note that $OPT_{\mathcal{K}_0}$ rounded up gives the exact value OPT_{MC} in almost all cases.

To explore the potential of our approach, we also generated some bigger instances with up to 100 nodes. We generated random graphs with edge probability p (g50.1, g50.4, g100) and also random graphs where the entries in the (1,2) block of size (m_1, m_2) are chosen with probability $q < p$. This should result in "easier" instances, as the partition $S_1 = \{1, \ldots, m_1\}$, $S_2 = \{m_1+1, \ldots, m_1+m_2\}$ has a smaller expected

TABLE 2
Some random graphs on n nodes.

Graph	n	$\|E\|$	p	q	m_1	m_2
g50.1	50	247	0.2			
g50.2	50	237	0.2	0.15	20	20
g50.3	50	198	0.2	0.10	25	20
g50.4	50	114	0.1			
g100	100	1000	0.2			

TABLE 3
Min-Cut approximation for the graphs from Table 2.

Graph	m_1, m_2, m_3	ubd	OPT_{HW}	$OPT_{\mathcal{K}_0^a}$	$m_3 + \lceil\sqrt{2\alpha}\rceil - 1$
g50.1	20 25 5	49	16.4	41.8	14
	20 20 10	29	-4.3	22.9	16
	15 20 15	14	-20.9	8.3	19
g50.2	20 25 5	45	17.2	36.7	13
	20 20 10	26	-4.9	19.1	16
	15 20 15	12	-22.6	5.8	18
g50.3	20 25 5	29	-1.1	24.1	12
	20 20 10	15	-17.5	10.1	14
	15 20 15	12	-22.6	5.8	16
g50.4	20 25 5	43	7.2	35.1	8
g100	40 50 10	251	161.7	225.1	31
	40 40 20	173	80.2	147.4	37
	35 35 30	107	12.5	84.7	43

number of edges (g50.2, g50.3). In Table 2 we provide some specifics about these graphs.

For these graphs the computation of the relaxation over $\mathcal{K}_0$ is beyond the possibilities of our computing facilities. The simpler relaxation over $\mathcal{K}_0^a$ can still be calculated rather easily. In Table 3 we summarize the results. The column labeled ubd gives an upper bound on OPT_{MC}, obtained by a simulated annealing heuristic. Then we compare $OPT_{\mathcal{K}_0^a}$ to the spectral bound OPT_{HW}. We partition the graphs in several ways, indicated by the vector m. We note also here that the new relaxation provides a substantial improvement over the original spectral bound, which in case of negative values does not give any relevant information at all. Further, more detailed, computational experiments will be reported elsewhere; see the dissertation [16].

6. Advances to the bandwidth and the vertex separator problem. For a graph G on n vertices we define a labeling of vertices as a bijection $\Phi\colon V = \{v_1, \ldots, v_n\} \to \{1, 2, \ldots, n\}$. The labeling bandwidth $\sigma_\infty(G, \Phi)$ of the labeling Φ is the maximal difference over all graph edges:

$$\sigma_\infty(G, \Phi) := \max_{(i,j)\in E} |\Phi(v_i) - \Phi(v_j)|.$$

The bandwidth of a graph G is the minimum of the labeling bandwidth over all labelings:

$$\sigma_\infty(G) := \min_{\Phi} \sigma_\infty(G, \Phi).$$

The bandwidth problem is an NP-hard problem and remains NP-hard even if the graph G is a tree with maximal degree at most 3 or a caterpillar with hairlength ≤ 3. Even approximating the bandwidth is an extremely difficult task. Blache et al. have

shown that there is no polynomial time algorithm with an approximation ratio smaller than 1.5 unless P = NP (for more results about the bandwidth problem and its complexity see [3, 4, 5, 12]). In [10] several lower bounds for σ_∞ have been established for an unweighted graph, using Laplacian eigenvalues of the graph. The basic tool the authors used was showing that $OPT_{MC} > 0$. If this is the case, then $\sigma_\infty(G) \geq m_3 + 1$. This is generalized in the following proposition.

PROPOSITION 12. *Let G be an undirected and unweighted graph. If for some $m = (m_1, m_2, m_3)$ it holds that $OPT_{MC} \geq \alpha > 0$, then*

$$\sigma_\infty(G) \geq \max\{m_3 + 1,\, m_3 + \lceil\sqrt{2\alpha}\rceil - 1\}.$$

Proof. Let Φ be the optimal labeling of G. We may assume that the vertices of G are initially labeled such that Φ is identity, i.e., $\Phi(i) = i$. Let (S_1, S_2, S_3) be a partition of $V(G)$, defined by $S_1 = \{1, \ldots, m_1\}$ and $S_2 = \{m_1 + m_3 + 1, \ldots, n\}$, Δ the maximal difference of end numbers over all edges, connecting sets S_1 and S_2, and $\delta = \Delta - m_3$. We have $\delta \geq 1$ since $OPT_{MC} > 0$. The only vertices from S_1 that might have a neighbor in S_2 are $m_1 - \delta + 1, \ldots, m_1$, since otherwise the difference of end vertices is greater than Δ. The same argument implies that the vertex m_1 has δ neighbors at most in S_2 , the vertex $m_1 - 1$ has $\delta - 1$ neighbors at most, etc. The last vertex $m_1 - \delta + 1$ has 1 neighbor at most in S_2. The number of edges between S_1 and S_2 is therefore $\delta + (\delta - 1) + \cdots + 1 = \delta(\delta + 1)/2$ at most; hence we get the inequality $\delta(\delta + 1) \geq 2\alpha$, which implies $\delta \geq \lceil\sqrt{2\alpha}\rceil - 1$. Since we also know that $\delta \geq 1$, the proposition follows from $\sigma_\infty(G) \geq \Delta$. □

Table 4 demonstrates the tightness of this lower bound on the graph instances from Table 1. The third column contains the bandwidth of the graph (for graphs $P_m \times P_n$ and $K_{m,n}$ we can compute it using the closed form formula, e.g., $\sigma_\infty(P_m \times P_n) = \min\{m, n\}$). In the fourth column we have α, the lower bound for OPT_{MC}, obtained by rounding up the best $OPT_{\mathcal{K}_0}$ from Table 1, and the last column shows the lower bound for $\sigma_\infty(G)$ from Proposition 12. We can see that we might get good information about the bandwidth using a good lower bound for the OPT_{MC}, and this is very important according to the complexity hardness of the bandwidth problem.

TABLE 4
Lower bounds for bandwidth for the graphs from Table 1.

Graph	m_3	$\sigma_\infty(G)$	α	$m_3 + \lceil\sqrt{2\alpha}\rceil - 1$
$P_6 \times P_4$	3	4	1	4
$P_6 \times P_4$	2	4	2	3
rand(15, 0.5)	4	10	7	7
$K_{6,9}$	5	10	5	8
$K_{6,9}$	4	10	9	8

Finally, we also compare the lower bound on the bandwidth from [10], which is based on the spectral bound OPT_{HW}, and the bound from Proposition 12 on the graphs from Table 2. The results in Table 5 show again that the new model provides a significant improvement over the spectral bound.

A set $S_3 \subset V$ is a vertex separator if removing these vertices disconnects the graph. It is a balanced vertex separator if the resulting graph has two components of sizes between $s/3$ and $2s/3$, where $s = |V| - |S_3|$. Helmberg et al. have derived in [10] several lower bounds on the size of a minimal vertex separator. They have used the fact that if $OPT_{MC} = 0$ then $OPT_{HW} \leq 0$ and from this have derived lower bounds on the size of the vertex separator. By using the fact that for fixed m_3 is OPT_{HW}

TABLE 5
Lower bounds on the bandwidth for the graphs from Table 2.

Graph	Bound from [10]	Proposition 12
g50.1	9	19
g50.2	9	18
g50.3	5	16
g50.4	5	8
g100	33	43

maximal if m_1 and m_2 are equal (or differs by 1 if $n - m_3$ is an odd number) they have extended the result to balanced vertex separators. The optimal values $OPT_{\mathcal{K}}$ for $\mathcal{K}$ as above give information about the vertex separators only if they are positive, since in this case we know that the graph does not have a vertex separator of size m_3 whose removal divides the graph vertices into sets of sizes m_1 and m_2. Table 1 shows that on the test instances we always detected the nonexistence of the appropriate vertex separator. However, since in general the value $OPT_{\mathcal{K}}$ does not monotonically change with the difference $|m_1 - m_2|$ as is the case for OPT_{HW}, we can get the information about the balanced vertex separator only by checking all possible pairs $m_2/2 \leq m_1 \leq m_2$ with $m_1 + m_2 = n - m_3$. This might be time consuming so it is worth trying to change the model MC_{CP} in order to include the balanced cardinality constraint and then relaxing this model. We have already done some promising steps and the results appear in [16].

7. Conclusions. We have shown that the MCP can be formulated as a linear program over the cone of completely positive matrices of order $3n \times 3n$. Replacing the cone of completely positive matrices with any cone for which we are able to solve the separation problem gives a tractable approximate model. We have analyzed the relaxation, obtained by using the cone of positive semidefinite matrices, and we showed that this model gives the eigenvalue lower bound, originally found by Helmberg et al. in [10]. We provided the closed form solution of this relaxation and showed that it is determined with the second and the largest eigenvalues of graph Laplacian and corresponding eigenvectors. We also proposed some other relaxations, using the hierarchy of cones, proposed by Parrilo in [15]. Numerical results in section 5 show that the lower bounds, obtained this way, may be very tight. At this point we want to emphasize that our approach may be easily extended to a general graph partitioning problem. We finished with the study of the impact that the new results have on approximation of some other combinatorial problems. A reasonable good lower bound for the bandwidth problem may be obtained this way as well as the certificate that the graph does not have a separator of specified size, whose removal disconnects the graph into two sets of prescribed sizes. A preliminary study in modeling the balanced vertex separator problem by copositive programming has also been done and the results together with an extension to the general graph partitioning problem will be reported elsewhere; see also the dissertation [16].

Acknowledgments. We thank two anonymous referees for several suggestions to improve an earlier version of the paper.

REFERENCES

[1] C. J. ALPERT AND A. B. KAHNG, *Recent directions in netlist partition: A survey*, Integration, the VLSI Journal, 19 (1995), pp. 1–81.

[2] K. Anstreicher and H. Wolkowicz, *On Lagrangian relaxation of quadratic matrix constraints*, SIAM J. Matrix Anal. Appl., 22 (2000), pp. 41–55.
[3] G. Blache, M. Karpinski, and W. Jürgen, *On approximation intractability of the bandwidth problem*, in Electronic Colloquium on Computation Complexity, University of Trier, Trier, Germany, 1998, TR98-014.
[4] F. R. K. Chung, *Labelings of graphs*, in Selected Topics in Graph Theory 3, Academic Press, San Diego, 1988, pp. 151–168.
[5] J. Díaz, J. Petit, and M. Serna, *A survey on graph layout problems*, ACM Comput. Surveys, 34 (2002), pp. 313–356.
[6] U. Feige, M. Hajiaghayi, and J. R. Lee, *Improved approximation algorithms for minimum-weight vertex separators*, in Proceedings of the 37th Annual ACM Symposium on Theory of Computing (STOC), Baltimore, MD, 2005, pp. 563–572.
[7] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, Freeman, San Francisco, 1979.
[8] S. Guattery and G. L. Miller, *On the quality of spectral separators*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 701–719.
[9] A. J. Hoffman and H. W. Wielandt, *The variation of the spectrum of a normal matrix*, Duke Math. J., 20 (1953), pp. 37–39.
[10] C. Helmberg, F. Rendl, B. Mohar, and S. Poljak, *A spectral approach to bandwidth and separator problems in graphs*, Linear and Multilinear Algebra, 39 (1995), pp. 73–90.
[11] E. de Klerk and D. V. Pasechnik, *Approximation of the stability number of a graph via copositive programming*, SIAM J. Optim., 12 (2002), pp. 875–892.
[12] Y.-L. Lai and K. Williams, *A survey of solved problems and applications on bandwidth, edgesum and profiles of graphs*, J. Graph Theory, 31 (1999), pp. 75–94.
[13] T. Lengauer, *Combinatorial Algorithms for Integrated Circuit Layout*, Wiley, Chichester, UK, 1990.
[14] K. G. Murty and S. N. Kabadi, *Some NP-complete problems in quadratic and nonlinear programming*, Math. Programming, 39 (1987), pp. 117–129.
[15] P. A. Parrilo, *Structured Semidefinite Programs and Semialgebraic Geometry Methods in Robustness and Optimization*, Ph.D. thesis, California Institute of Technology, Pasadena, CA, 2000.
[16] J. Povh, *Application of Semidefinite and Copositive Programming in Combinatorial Optimization*, Ph.D. thesis, University of Ljubljana, Ljubljana, Slovenia, 2006.
[17] F. Rendl and H. Wolkowicz, *A projection technique for partitioning the nodes of a graph*, Ann. Oper. Res., 58 (1995), pp. 155–179.
[18] C. de Sousa and E. Balas, *The vertex separator problem, algorithms and computations*, Math. Program., 103 (2005), pp. 609–631.
[19] H. Wolkowicz and Q. Zhao, *Semidefinite programming relaxations for the graph partitioning problem*, Discrete Appl. Math., 96–97 (1999), pp. 461–479.

SIAM J. OPTIM.
Vol. 18, No. 1, pp. 242–259

A BUNDLE METHOD FOR A CLASS OF BILEVEL NONSMOOTH CONVEX MINIMIZATION PROBLEMS*

MIKHAIL V. SOLODOV†

Abstract. We consider the bilevel problem of minimizing a nonsmooth convex function over the set of minimizers of another nonsmooth convex function. Standard convex constrained optimization is a particular case in this framework, corresponding to taking the lower level function as a penalty of the feasible set. We develop an explicit bundle-type algorithm for solving the bilevel problem, where each iteration consists of making one descent step for a weighted sum of the upper and lower level functions, after which the weight can be updated immediately. Convergence is shown under very mild assumptions. We note that in the case of standard constrained optimization, the method does not require iterative solution of any penalization subproblems—not even approximately—and does not assume any regularity of constraints (e.g., the Slater condition). We also present some computational experiments for minimizing a nonsmooth convex function over a set defined by linear complementarity constraints.

Key words. bilevel optimization, convex optimization, nonsmooth optimization, bundle methods, penalty methods

AMS subject classifications. 90C30, 65K05, 49D27

DOI. 10.1137/050647566

1. Introduction.

We consider a class of *bilevel problems* of the form

$$\begin{array}{ll} \text{minimize} & f_1(x) \\ \text{subject to} & x \in S_2 = \arg\min\{f_2(x) \mid x \in \Re^n\}, \end{array} \tag{1.1}$$

where $f_1 : \Re^n \to \Re$ and $f_2 : \Re^n \to \Re$ are convex functions, in general nondifferentiable.

The above is a special case of the *mathematical program with generalized equation (or equilibrium) constraint* [20, 7], which is

$$\begin{array}{ll} \text{minimize} & f_1(x) \\ \text{subject to} & x \in \{x \in \Re^n \mid 0 \in T(x)\}, \end{array}$$

where T is a set-valued mapping from $\Re^n$ to the subsets of $\Re^n$. The bilevel problem (1.1) is obtained by setting $T(x) = \partial f_2(x)$, $x \in \Re^n$. In the formulation of the problem considered here, there is only one (decision) variable $x \in \Re^n$, and we are interested in identifying specific solutions of the inclusion $0 \in T(x)$ (equivalently, of the lower level minimization problem in (1.1)); see [7]. Problems of the form of (1.1) are also sometimes referred to as *hierarchical optimization*; see, e.g., [12, 4].

Note that, as a special case, (1.1) contains the standard convex constrained optimization problem

$$\begin{array}{ll} \text{minimize} & f_1(x) \\ \text{subject to} & g_i(x) \le 0, \quad i = 1, \ldots, m, \end{array} \tag{1.2}$$

*Received by the editors December 14, 2005; accepted for publication (in revised form) October 24, 2006; published electronically April 3, 2007. This research is supported in part by CNPq grants 301508/2005-4, 490200/2005-2, and 550317/2005-8, by PRONEX-Optimization, and by FAPERJ grant E-26/151.942/2004.

http://www.siam.org/journals/siopt/18-1/64756.html

†Instituto de Matemática Pura e Aplicada, Estrada Dona Castorina 110, Jardim Botânico, Rio de Janeiro, RJ 22460-320, Brazil (solodov@impa.br).

where $g : \Re^n \to \Re^m$ is a (nonsmooth) convex function. Indeed, (1.2) is obtained from (1.1) by taking $f_2(x) = p(x)$, where $p : \Re^n \to \Re_+$ is some penalty of the constraints, e.g.,

$$f_2(x) = p(x) = \sum_{i=1}^{m} \max\{0,\, g_i(x)\}. \tag{1.3}$$

In this paper, we show that the bilevel problem (1.1) can be solved by a properly designed (proximal) bundle method [15, 11, 3], iteratively applied to the parametrized family of functions

$$F_\sigma(x) = \sigma f_1(x) + f_2(x), \quad \sigma > 0, \tag{1.4}$$

where σ varies along the iterations. Specifically, if $x^k \in \Re^n$ is the current iterate and $\sigma_k > 0$ is the current parameter, it is enough to make *just one* descent step for F_{σ_k} from the point x^k, after which the parameter σ_k can be immediately updated. We emphasize that at no iteration is the function F_{σ_k} minimized to any prescribed precision. Once the descent condition is achieved, the parameter can be updated immediately and we can start working with the new function $F_{\sigma_{k+1}}$. For convergence of the resulting algorithm to the solution set of (1.1), parameters $\{\sigma_k\}$ should be chosen is such a way that

$$\lim_{k\to\infty} \sigma_k = 0, \quad \sum_{k=0}^{\infty} \sigma_k = +\infty. \tag{1.5}$$

The requirement that σ_k must tend to zero is natural and indispensable, as can be seen from the case of standard optimization (1.2). To this end, it is interesting to comment on the relation between our method and the classical penalty approximation scheme [8, 23]. The penalty scheme consists of solving a sequence of unconstrained subproblems

$$\text{minimize } F_\sigma(x), \quad x \in \Re^n, \tag{1.6}$$

where F_σ is given by (1.4) with f_2 being a penalty term p, such as (1.3). (In the literature, it is more common to minimize $\sigma^{-1}F_\sigma(x) = f_1(x) + \sigma^{-1}p(x)$, but the resulting subproblem is clearly equivalent to (1.6).) As is well known, under mild assumptions optimal paths of solutions $x(\sigma)$ of penalized problems (1.6) tend to the solution set of (1.2) as $\sigma \to 0$. We emphasize that the requirement that penalty parameters should tend to zero is, in general, indispensable. To guarantee that a solution of (1.6) is a solution of the original problem (1.2) for some *fixed* $\sigma > 0$ (i.e., exactness of the penalty function), some regularity assumptions on constraints are needed (e.g., see [3, section 14.4]). No assumptions of this type are made in this paper. The fundamental issue is approximating $x(\sigma_k)$ for some sequence of parameters $\sigma_k \to 0$. It is clear that approximating $x(\sigma_k)$ with precision is computationally impractical. It is therefore attractive to trace the optimal path in a loose (and computationally cheap) manner, while still safeguarding convergence. In a sense, this is what our method does: instead of solving subproblems (1.6) to some prescribed accuracy, it makes just one descent step for F_{σ_k} from the current iterate x^k and immediately updates the parameter. We emphasize that this results in meaningful progress (and ultimately produces iterates converging to solutions of the problem) for arbitrary points x^k, and not just for points close to the optimal path, i.e., points close to $x(\sigma_k)$.

We therefore obtain an implementable algorithm for tracing optimal paths of penalty schemes.

We next discuss the relationship of our algorithm to the existing literature. For the bilevel setting of (1.1), we believe that our proposal is the first method which is completely *explicit*. In some ways, it is related to [4], where a proximal point method for (1.1) has been considered, and (1.5) is referred to as *slow control*. However, as any proximal method, the method of [4] is *implicit*: it requires solving nontrivial subproblems of minimizing regularizations of functions F_{σ_k} at every iteration, even if approximately. By contrast, the method proposed in this paper is completely explicit: each iteration is a serious (or descent) step for the current F_{σ_k}, constructed by a finite number of null steps in a way which is essentially standard in nonsmooth optimization.

The special case of standard optimization deserves some further comments. We next discuss bundle methods applicable to problems with nonlinear constraints, such as (1.2) above. When the problem admits exact penalization, one can solve the equivalent unconstrained problem of minimizing the exact penalty function; see [14, 18]. However, as already mentioned above, exact penalization requires regularity assumptions on constraints, such as the Slater condition (existence of some $x \in \Re^n$ such that $g_i(x) < 0$ for all $i = 1, \ldots, m$). We stress that no assumptions of this type are needed for our method. For example, our method is applicable to minimizing a nonsmooth function subject to (monotone linear) complementarity constraints

$$(Qx + q)_i \geq 0, \quad x_i \geq 0, \quad x_i(Qx + q)_i = 0, \quad i = 1, \ldots, n,$$

where Q is an $n \times n$ positive semidefinite matrix. Those constraints can be modeled in the form (1.2) as

$$-Qx - q \leq 0, \quad -x \leq 0, \quad \langle Qx + q, x \rangle \leq 0.$$

Complementarity constraints do not satisfy constraint qualifications, no matter how they are modeled, which makes this class of problems particularly difficult. We shall come back to problems with complementarity constraints in section 4, where some computational experiments are presented.

The methods in [21, 22] and [15, Chap. 5] do not use penalization but enforce feasibility of every serious iteration. In particular, they require a feasible starting point, which is a difficult computational task (in the case of nonlinear constraints). In addition, regularity of constraints is still needed for convergence. Bundle methods which do not use penalty functions and do not enforce feasibility are [19, 9, 24, 13]. The methods in [24, 13] share one feature in common with the one proposed here: they apply bundle techniques to a dynamically changing objective function, except that the function is different (underlying [24, 13] is the so-called *improvement function*, which goes back to [21, 15, 1]). The methods of [24, 13] require the Slater condition, while those in [19, 9] do not. However, [19, 9] (as well as [21, 17, 18]) need a priori boundedness assumptions on the iterates to prove convergence. For our method, we assume only that the solution set of the problem is bounded.

For the standard optimization setting (1.2), this paper is also somewhat related to [5], where interior penalty schemes are coupled with continuous-time steepest descent to produce a family of paths converging to a solution set. However, concrete numerical schemes in [5] arise from *implicit* discretization and, thus, result in implicit proximal-point iterations, just as in [4]. Nevertheless, it was conjectured in [5] that an economic algorithm performing a single iteration of some descent method for each value of σ_k could be enough to generate a sequence of iterates converging to a solution of the

problem. This is what the presented method does, although we use exterior rather than interior penalties and consider the more general nonsmooth setting (as well as the more general bilevel setting). A related explicit descent scheme for the smooth case has been developed in [25].

Our notation is quite standard. By $\langle x, y\rangle$ we denote the inner product of x and y, and by $\|\cdot\|$ the associated norm, where the space is always clear from the context. For a convex function $f : \Re^n \to \Re$, its ε-subdifferential at the point $x \in \Re^n$ is denoted by $\partial_\varepsilon f(x) = \{g \in \Re^n \mid f(y) \geq f(x) + \langle g, y - x\rangle - \varepsilon \text{ for all } y \in \Re^n\}$, where $\varepsilon \in \Re_+$. Then the subdifferential of f at x is given by $\partial f(x) = \partial_0 f(x)$. If S is a closed convex set in $\Re^n$, then $P_S(x)$ stands for the orthogonal projection of the point $x \in \Re^n$ onto S, and $\operatorname{dist}(x, S) = \|x - P_S(x)\|$ is the distance from x to S.

2. The algorithm. As already outlined above, the conceptual idea of the algorithm is quite simple. If $x^k \in \Re^n$ is the current approximation to a solution of (1.1) and $\sigma_k > 0$ is the current parameter defining the function F_{σ_k} in (1.4), an iteration of the method consists of making a descent step for F_{σ_k} relative to its value at x^k. After this, the value of σ_k can be changed immediately. Since the function F_{σ_k} is nonsmooth, the computationally implementable way to construct a descent step is the bundle technique [15, 11, 3]. We next introduce the notation necessary for stating our algorithm.

Bundle methods keep memory of the past in a *bundle* of information. Let x^k be the current approximation to a solution and let y^i, $i = 1, \ldots, \ell - 1$, be all the points that have been produced by the method so far, including the ones which have not been accepted as satisfactory (so-called "null steps"). Generally, $\{x^k\}$ is a particular subsequence of $\{y^i\}$. For an iteration index ℓ, we shall denote by $k(\ell)$ the index of the last iteration preceding the iteration ℓ at which x^k and σ_k have been modified. Whenever k and ℓ appear in the same expression, we mean that $k = k(\ell)$.

Let us denote the function and subgradient values of f_1 at the points y^i, $i = 1, \ldots, \ell - 1$, by $f_1^i = f_1(y^i)$, $g_1^i \in \partial f_1(y^i)$, and similarly for f_2. Since $(\sigma_k g_1^i + g_2^i) \in \partial F_{\sigma_k}(y^i)$, this information can be used to define a cutting-planes approximation Ψ_ℓ of the function F_{σ_k}, as follows:

$$\begin{aligned}\Psi_\ell(y) &:= \max_{i<\ell} \left\{\sigma_k f_1^i + f_2^i + \langle \sigma_k g_1^i + g_2^i, y - y^i\rangle\right\} \\ &= \sigma_k f_1(x^k) + f_2(x^k) \\ &\quad + \max_{i<\ell} \left\{-(\sigma_k e_1^{k,i} + e_2^{k,i}) + \langle \sigma_k g_1^i + g_2^i, y - x^k\rangle\right\}, \end{aligned} \tag{2.1}$$

where the second expression is centered at x^k and uses the linearization errors at y^i with respect to x^k:

$$e_p^{k,i} := f_p(x^k) - f_p^i - \langle g_p^i, x^k - y^i\rangle \geq 0, \quad p = 1, 2. \tag{2.2}$$

We note that the second representation of Ψ_ℓ in (2.1) is better suited for implementations, due to lower storage requirements. As is readily seen from the definition of ε-subgradients, it holds that

$$g_p^i \in \partial_{e_p^{k,i}} f_p(x^k), \quad p = 1, 2, \tag{2.3}$$

and

$$(\sigma_k g_1^i + g_2^i) \in \partial_{(\sigma_k e_1^{k,i} + e_2^{k,i})} F_{\sigma_k}(x^k). \tag{2.4}$$

The linearization errors in (2.2) have to be properly updated every time x^k changes.

Choosing a proximal parameter $\mu_\ell > 0$, we generate the next candidate point y^ℓ by solving a quadratic programming (QP) reformulation of the problem

$$\min_{y\in\Re^n}\left\{\Psi_\ell(y)+\frac{1}{2}\mu_\ell\|y-x^k\|^2\right\}. \tag{2.5}$$

We note that the resulting quadratic program possesses a certain special structure, for which efficient software has been developed [16, 10]. The iterate y^ℓ is considered good enough when $F_{\sigma_k}(y^\ell)$ is sufficiently smaller than $F_{\sigma_k}(x^k)$ (the so-called "serious step"; this will be made precise later). If y^ℓ is acceptable, then we set $x^{k+1} := y^\ell$, choose new σ_{k+1}, and proceed to construct a descent step for $F_{\sigma_{k+1}}$. Otherwise, a so-called "null step" is declared and the procedure continues for F_{σ_k}, using the enhanced approximation $\Psi_{\ell+1}$.

In order for the basic idea outlined above to be practical, some important details have to be incorporated into the design of the method, as discussed next.

The number of constraints in the QP reformulation of (2.5) is precisely the number of elements in the bundle. Obviously, one has to keep this number computationally manageable. Thus, the bundle has to be *compressed* whenever the number of elements reaches some chosen bound. Reducing the bundle amounts to replacing the cutting-planes model (2.1) with another function, defined with a smaller number of cutting planes, which we shall still denote by Ψ_ℓ. This has to be done without impairing convergence of the algorithm. For this purpose, the so-called *aggregate function* is fundamental [3, Chap. 9], which we shall introduce in what follows.

It is convenient to split the information kept at iteration ℓ into two separate parts. One is the "oracle" bundle containing subgradient values at (some of!) points y^i, $i = 1, \ldots, \ell-1$, and the associated linearization errors (recall (2.3) and (2.2)):

$$\mathcal{B}_\ell^{oracle} \subset \bigcup_{i<\ell}\left\{\left(e_p^{k,i}\in\Re_+, g_p^i\in\partial_{e_p^{k,i}} f_p(x^k),\ p=1,2\right)\right\}.$$

Note that here, the bundle $\mathcal{B}_\ell^{oracle}$ is not required to contain information at all the previous points (this is reflected by the use of the inclusion, rather than equation, in the definition above). The other part is the "aggregate" bundle, obtained from solutions of the QP subproblems. This bundle contains certain special ε-subgradients at x^k, to be introduced in Lemma 2.1 below. For now, we formally set

$$\mathcal{B}_\ell^{agg} \subset \bigcup_{i<\ell}\left\{\left(\hat\varepsilon_p^{k,i}\in\Re_+, \hat g_p^i\in\partial_{\hat\varepsilon_p^{k,i}} f_p(x^k),\ p=1,2\right)\right\},$$

without specifying how exactly those objects are obtained. Note that here there may no longer exist any previous point y^i, $i<\ell$, for which $\hat g_p^i \in \partial f_p(y^i)$, $p = 1, 2$.

The information in $\mathcal{B}_\ell^{oracle}$ and $\mathcal{B}_\ell^{agg}$ defines a cutting-planes approximation of F_{σ_k} given by

$$\begin{aligned}\Psi_\ell(y) = {}& \sigma_k f_1(x^k) + f_2(x^k)\\ &+\max\left\{\max_{i\in\mathcal{B}_\ell^{oracle}}\left\{-(\sigma_k e_1^{k,i}+e_2^{k,i})+\langle\sigma_k g_1^i+g_2^i, y-x^k\rangle\right\},\right.\\ &\left.\max_{i\in\mathcal{B}_\ell^{agg}}\left\{-(\sigma_k\hat\varepsilon_1^{k,i}+\hat\varepsilon_2^{k,i})+\langle\sigma_k\hat g_1^i+\hat g_2^i, y-x^k\rangle\right\}\right\},\end{aligned} \tag{2.6}$$

where by $i \in \mathcal{B}_\ell^{oracle}$ we mean that there exists an element in the set $\mathcal{B}_\ell^{oracle}$ indexed by i; and similarly for $\mathcal{B}_\ell^{agg}$. Although this notation is formally improper (the bundles are not sets of indices), it does not lead to any confusion while simplifying the formulas.

We next discuss properties of the solution of QP subproblem (2.5) with Ψ_ℓ given by (2.6). The following characterization is an adaptation of [3, Lemma 9.8] for our setting.

LEMMA 2.1. *For the unique solution y^ℓ of (2.5) with Ψ_ℓ given by (2.6), it holds that*

(i) $y^\ell = x^k - \frac{1}{\mu_\ell}(\sigma_k \hat{g}_1^\ell + \hat{g}_2^\ell)$;
(ii) $\hat{g}_p^\ell = \sum_{i \in \mathcal{B}_\ell^{oracle}} \lambda_i^\ell g_p^i + \sum_{i \in \mathcal{B}_\ell^{agg}} \hat{\lambda}_i^\ell \hat{g}_p^i$, $p = 1, 2$,
where $\lambda^\ell \geq 0$, $\hat{\lambda}^\ell \geq 0$ and $\sum_{i \in \mathcal{B}_\ell^{oracle}} \lambda_i^\ell + \sum_{i \in \mathcal{B}_\ell^{agg}} \hat{\lambda}_i^\ell = 1$;
(iii) $(\sigma_k \hat{g}_1^\ell + \hat{g}_2^\ell) \in \partial \Psi_\ell(y^\ell)$;
(iv) $\hat{g}_p^\ell \in \partial_{\hat{\varepsilon}_p^{k,\ell}} f_p(x^k)$, *where* $\hat{\varepsilon}_p^{k,\ell} = \sum_{i \in \mathcal{B}_\ell^{oracle}} \lambda_i^\ell e_p^{k,i} + \sum_{i \in \mathcal{B}_\ell^{agg}} \hat{\lambda}_i^\ell \hat{\varepsilon}_p^{k,i}$, $p = 1, 2$;
(v) $(\sigma_k \hat{g}_1^\ell + \hat{g}_2^\ell) = \hat{g}^\ell \in \partial_{\hat{\varepsilon}^{k,\ell}} F_{\sigma_k}(x^k)$, *where* $\hat{\varepsilon}^{k,\ell} = \sigma_k \hat{\varepsilon}_1^{k,\ell} + \hat{\varepsilon}_2^{k,\ell}$;
(vi) $\hat{\varepsilon}^{k,\ell} = F_{\sigma_k}(x^k) - \Psi_\ell(y^\ell) - \frac{1}{\mu_\ell}\|\hat{g}^\ell\|^2 \geq 0$.

Proof. The assertions can be verified following the analysis in [3, Lemma 9.8], and taking into account the special structure of the function F_{σ_k} and of its approximation Ψ_ℓ. We omit the details. □

We note that λ^ℓ and $\hat{\lambda}^\ell$ in Lemma 2.1 are the Lagrange multipliers associated with y^ℓ in the quadratic program reformulation of (2.5) (or the problem variables, if one solves the dual of this quadratic program, as in [16]). In any case, λ^ℓ and $\hat{\lambda}^\ell$ are available as part of the solution to (2.5). The quantities $\hat{g}_p^\ell$, $\hat{\varepsilon}_p^{k,\ell}$, $p = 1, 2$, defined in Lemma 2.1 are precisely the ones that appear in the definition of $\mathcal{B}_\ell^{agg}$ (except that $\mathcal{B}_\ell^{agg}$ contains information computed at iterations previous to the ℓth; at the first iteration we formally set $\mathcal{B}_\ell^{agg} = \emptyset$). We are now ready to introduce the aggregate function, already mentioned above:

$$l_{k,\ell}(y) := \sigma_k f_1(x^k) + f_2(x^k) - (\sigma_k \hat{\varepsilon}_1^{k,\ell} + \hat{\varepsilon}_2^{k,\ell}) + \langle \sigma_k \hat{g}_1^\ell + \hat{g}_2^\ell, y - x^k \rangle,$$

where

$$\hat{g}_p^\ell \in \partial_{\hat{\varepsilon}_p^{k,\ell}} f_p(x^k), \quad p = 1, 2, \tag{2.7}$$

and consequently,

$$(\sigma_k \hat{g}_1^\ell + \hat{g}_2^\ell) \in \partial_{(\sigma_k \hat{\varepsilon}_1^{k,\ell} + \hat{\varepsilon}_2^{k,\ell})} F_{\sigma_k}(x^k). \tag{2.8}$$

As already noted above, this function is defined directly from the quantities available after solving (2.5).

As pointed out in [6, eqs. (4.7)–(4.9)], to guarantee that a bundle technique would be able to construct a descent step for F_{σ_k} with respect to its value at x^k (assuming x^k is not a minimizer of F_{σ_k}) one can actually use any cutting-planes models Ψ_ℓ satisfying (for all $y \in \Re^n$) the following three conditions:

$$\begin{aligned} \Psi_\ell(y) &\leq F_{\sigma_k}(y) && \text{for all } \ell \geq 1 \text{ and all } k, \\ l_{k,\ell}(y) &\leq \Psi_{\ell+1}(y) && \text{for those } \ell \text{ for which } y^\ell \text{ is a null step}, \\ \sigma_k f_1^\ell + f_2^\ell + \langle \sigma_k g_1^\ell + g_2^\ell, y - y^\ell \rangle &\leq \Psi_{\ell+1}(y) && \text{for those } \ell \text{ for which } y^\ell \text{ is a null step}. \end{aligned}$$

The last two conditions mean that when defining the new bundles, it is enough for $\mathcal{B}_{\ell+1}^{oracle}$ to contain the cutting plane computed at the new point y^ℓ (i.e., the subgradients g_1^ℓ, g_2^ℓ, and the associated linearization errors $e_1^{k,\ell}$, $e_2^{k,\ell}$) and for $\mathcal{B}_{\ell+1}^{agg}$ to contain

the last aggregate function $l_{k,\ell}$ (i.e., the ε-subgradients $\hat{g}_1^\ell$, $\hat{g}_2^\ell$, and the associated $\hat{\varepsilon}_1^{k,\ell}$, $\hat{\varepsilon}_2^{k,\ell}$). In particular, at any iteration, the bundle can contain as few elements as we wish (as long as the two specified above are included). This fact is crucial for effective control of the size of subproblems (2.5). Finally, to make sure that the first condition above holds for all k, the linearization and aggregate errors have to be properly updated every time x^k changes to x^{k+1} (in particular, to ensure the key relations (2.4) and (2.8)). As is readily seen, the following formulas do the job:

(2.9)
$$e_p^{k+1,i} = e_p^{k,i} + f_p(x^{k+1}) - f_p(x^k) + \langle g_p^i, x^k - x^{k+1}\rangle, \quad p = 1,2, \text{ for } i \in \mathcal{B}_{\ell+1}^{oracle},$$
$$\hat{\varepsilon}_p^{k+1,i} = \hat{\varepsilon}_p^{k,i} + f_p(x^{k+1}) - f_p(x^k) + \langle \hat{g}_p^i, x^k - x^{k+1}\rangle, \quad p = 1,2, \text{ for } i \in \mathcal{B}_{\ell+1}^{agg}.$$

We are now ready to formally state the algorithm.

ALGORITHM 2.1 (bilevel bundle method).

Step 0. `Initialization.`
Choose parameter $m \in (0,1)$ *and an integer* $|\mathcal{B}|_{max} \geq 2$.
Choose $x^0 \in \Re^n$ *and* $\sigma_0 > 0$, $\beta_0 > 0$. *Set* $y^0 := x^0$ *and compute* f_p^0, g_p^0, $p = 1,2$. *Set* $k = 0$, $\ell = 1$, $e_p^{0,0} := 0$, $p = 1,2$. *Define the starting bundles* $\mathcal{B}_1^{oracle} := \{(e_p^{0,0}, g_p^0,\ p = 1,2)\}$ *and* $\mathcal{B}_1^{agg} := \emptyset$.

Step 1. `QP subproblem.`
Choose $\mu_\ell > 0$ *and compute* y^ℓ *as the solution of* (2.5), *where* Ψ_ℓ *is defined by* (2.6). *Compute*

$$\hat{g}^\ell = \mu_\ell(x^k - y^\ell), \quad \hat{\varepsilon}^{k,\ell} = F_{\sigma_k}(x^k) - \Psi_\ell(y^\ell) - \frac{1}{\mu_\ell}\|\hat{g}^\ell\|^2, \quad \delta_\ell = \hat{\varepsilon}^{k,\ell} + \frac{1}{2\mu_\ell}\|\hat{g}^\ell\|^2.$$

Compute f_p^ℓ, g_p^ℓ, $p = 1,2$. *Compute* $e_p^{k,\ell}$, $p = 1,2$, *using* (2.2) *written with* $i = \ell$.

Step 2. `Descent test.` *If*

$$F_{\sigma_k}(y^\ell) \leq F_{\sigma_k}(x^k) - m\delta_\ell, \tag{2.10}$$

then declare a serious step. *Otherwise, declare a* null step.

Step 3. `Bundle management.`
Set $\mathcal{B}_{\ell+1}^{oracle} := \mathcal{B}_\ell^{oracle}$ *and* $\mathcal{B}_{\ell+1}^{agg} := \mathcal{B}_\ell^{agg}$. *If the bundle has reached the maximum size (i.e., if* $|\mathcal{B}_{\ell+1}^{oracle} \cup \mathcal{B}_{\ell+1}^{agg}| = |\mathcal{B}|_{max}$*), then delete at least two elements from* $\mathcal{B}_{\ell+1}^{oracle} \cup \mathcal{B}_{\ell+1}^{agg}$ *and append the aggregate information* $(\hat{\varepsilon}_p^{\ell,k}, \hat{g}_p^\ell,\ p = 1,2)$ *to* $\mathcal{B}_{\ell+1}^{agg}$.
In any case, append $(e_p^{k,\ell}, g_p^\ell,\ p = 1,2)$ *to* $\mathcal{B}_{\ell+1}^{oracle}$.

Step 4. *If* `Descent test` *was satisfied,*
set $x^{k+1} = y^\ell$ *and choose* $0 < \sigma_{k+1} \leq \sigma_k$ *and* $0 < \beta_{k+1} \leq \beta_k$.
Update the linearization and aggregate errors using (2.9).
Set $k = k + 1$ *and go to Step* 5.
If

$$\max\{\hat{\varepsilon}^{k,\ell}, \|\hat{g}^\ell\|\} \leq \beta_k\sigma_k, \tag{2.11}$$

choose $0 < \sigma_{k+1} < \sigma_k$ *and* $0 < \beta_{k+1} < \beta_k$.
Set $x^{k+1} = x^k$, $k = k + 1$ *and go to Step* 5.

Step 5. *Set* $\ell = \ell + 1$ *and go to Step* 1.

The role of checking condition (2.11) is to detect the situation when the point x^k happens to be a minimizer of the function F_{σ_k} (or is almost a minimizer; recall Lemma 2.1(v)). If it is so, we immediately update the parameter σ_k. This is reasonable, since we are not interested in minimizing F_{σ_k}. The case of x^k being a minimizer of F_{σ_k}, however, is very unlikely to occur, since for no iteration k the function F_{σ_k} is being minimized with any prescribed precision. This is also confirmed by our numerical experiments in section 4, where we ignored the safeguard (2.11) in our implementation.

The algorithm does not have an overall stopping test. In the unconstrained case, a reliable stopping test is one of the important advantages of bundle methods (as compared, for example, to subgradient methods). However, lack of a stopping test in our setting cannot be considered to be a drawback of the algorithm. Indeed, a bilevel problem does not admit an explicit optimality condition. Actually, the same is in general already true for constrained optimization without a regularity assumption on the constraints (except for some special cases, of course). As a result, there is no explicit way to measure violation/satisfaction of optimality in (1.1), and, consequently, lack of a stopping test is inherent in the nature of the problem.

We note that there is certain freedom in updating or not updating the parameter σ_k after every iteration. While our goal is to show that we can update it after a single descent step, note that, in principle, we are not obliged to do so ($\sigma_{k+1} = \sigma_k$ is allowed, unless (2.11) holds; in the latter case, x^k almost minimizes F_{σ_k} and it does not make sense to insist on further descent for this function). For convergence, it would be required that σ_k not go to zero too fast, in the sense of condition (1.5) stated above. In the case of the standard optimization problem (1.2), this condition allows a natural interpretation. In order to be able to trace the optimal penalty path $x(\sigma)$ with such a relaxed precision (making just one descent step for each penalized subproblem (1.6)), we should not be jumping too far from the target $x(\sigma_k)$ on the path to the next target $x(\sigma_{k+1})$ as we move along. On the other hand, if σ_k is kept constant over a few descent iterations, this allows for a more rapid change in the parameter for the next iteration, while still guaranteeing the second condition in (1.5). This is intuitively reasonable: if we get closer to the optimal path, then the target can be moved further. In our numerical experiments in section 4, we have used the simplest generic choice of $\sigma_k = \sigma_0/(k+1)$. We have experimented with some other options (for example, keeping the parameter unchanged for some iterations), but found that this does not make much difference (for our test problems). We shall discuss this further in section 4.

3. Convergence analysis. In our convergence analysis, we assume that the objective function f_1 is bounded below; i.e.,

$$-\infty < \bar{f}_1 = \inf\{f_1(x) \mid x \in \Re^n\}.$$

Since we also assume that the problem is solvable, the function f_2 is automatically bounded below, and we define

$$-\infty < \bar{f}_2 = \min\{f_2(x) \mid x \in \Re^n\}.$$

For the subsequent analysis, it is convenient to think of Algorithm 2.1 as "applied" to the shifted function

$$F_\sigma(x) = \sigma(f_1(x) - \bar{f}_1) + (f_2(x) - \bar{f}_2), \tag{3.1}$$

instead of the function F_σ given by (1.4), as stated originally. We can do this because Algorithm 2.1 would generate the same iterates whether F_σ were given by (3.1) or

(1.4). Indeed, the two functions have the same subgradients and the same difference for function values at any two points. Hence, the cutting-planes models (2.6) for the two functions would differ by a constant term (not dependent on y). This means that solutions y^ℓ of QP subproblems (2.5) would be the same, as well as the quantities $\hat{g}^\ell$ and $\hat{\varepsilon}^{k,\ell}$, which are defined by those solutions. Therefore, the relations in (2.10) and (2.11), which are guiding the algorithm, also do not change. From now on, we consider that the method is "applied" to function F_σ defined by (3.1) (even though the function from (1.4) is used in reality, of course). This is convenient for the subsequent analysis and should not lead to any confusion.

We proceed to prove convergence of the algorithm.

PROPOSITION 3.1. *Let f_1 and f_2 be convex functions.*

If for consecutive null steps it holds that $\bar{\mu} \geq \mu_{\ell+1} \geq \mu_\ell > 0$, then Algorithm 2.1 *is well defined and either* (2.10) *or* (2.11) *(or both) hold infinitely often. In particular, the parameter σ_k is updated infinitely often.*

Proof. Let k be any iteration index and consider the sequence of null steps applied to the current (fixed over those null steps) function F_{σ_k}. By properties of standard bundle methods (e.g., [3, Thm. 9.15]), it holds that either the descent test (2.10) is satisfied after a finite number of null steps, or x^k is a minimizer of F_{σ_k}. In the latter case, it further holds that $\delta_\ell \to 0$ as $\ell \to \infty$. Hence, $\hat{g}^\ell \to 0$ and $\hat{\varepsilon}^{k,\ell} \to 0$ as $\ell \to \infty$. This means that the condition (2.11) would be satisfied after a finite number of null steps.

We have therefore established that either (2.10) or (2.11) is guaranteed to be satisfied after a finite number of null steps. This shows that the method is well defined and updates σ_k infinitely often. □

We next prove that the generated sequence $\{x^k\}$ is bounded and its accumulation points are feasible for problem (1.1).

PROPOSITION 3.2. *Let f_1 and f_2 be convex functions such that f_1 is bounded below on $\Re^n$ and the solution set S_1 of problem* (1.1) *is nonempty and bounded.*

Suppose that $\bar{\mu} \geq \mu_\ell \geq \hat{\mu} > 0$ for all iterations ℓ, that $\mu_{\ell+1} \geq \mu_\ell$ on consecutive null steps, and that $\sigma_k \to 0$ as $k \to \infty$.

Then any sequence $\{x^k\}$ generated by Algorithm 2.1 *is bounded and all its accumulation points are feasible for problem* (1.1); *i.e., they belong to S_2.*

Proof. If the serious step descent test (2.10) is satisfied only a finite number of times, it is readily seen that there exists some iteration index k_0 such that $x^k = x^{k_0}$ for all $k \geq k_0$ (because x^k is changed only at serious steps, i.e., when (2.10) holds). Hence, in this case $\{x^k\}$ is trivially bounded.

Assume now that (2.10) is satisfied infinitely often. In what follows, we consider the subsequence of indices k at which (2.10) holds, i.e., at which x^k changes. But to simplify the notation, we shall not introduce this subsequence explicitly. Here, we can simply disregard all the iterations at which x^k remained fixed. We can do this within the current analysis of boundedness of $\{x^k\}$, because those iterations merely changed σ_k (and the only assumption for the latter used below is that it should be nonincreasing—the property which holds for any subsequence of $\{\sigma_k\}$ by the construction of the method).

For each k, let $\ell(k)$ be the index ℓ for which (2.10) was satisfied (in particular, $x^{k+1} = y^{\ell(k)}$). By (2.10), it holds that

$$\begin{aligned} m\delta_{\ell(k)} &\leq F_{\sigma_k}(x^k) - F_{\sigma_k}(x^{k+1}) \\ &= \sigma_k(f_1(x^k) - \bar{f}_1) - \sigma_k(f_1(x^{k+1}) - \bar{f}_1) \\ &\quad + (f_2(x^k) - \bar{f}_2) - (f_2(x^{k+1}) - \bar{f}_2). \end{aligned}$$

Summing up the latter inequalities for $k = 0, \dots, k_1$, we obtain that

$$\begin{aligned} m\sum_{k=0}^{k_1}\delta_{\ell(k)} &\le \sigma_0(f_1(x^0)-\bar f_1) + \sum_{k=0}^{k_1-1}(\sigma_{k+1}-\sigma_k)(f_1(x^{k+1})-\bar f_1) \\ &\quad - \sigma_{k_1}(f_1(x^{k_1+1})-\bar f_1) + (f_2(x^0)-\bar f_2) - (f_2(x^{k_1+1})-\bar f_2) \\ &\le \sigma_0(f_1(x^0)-\bar f_1) + (f_2(x^0)-\bar f_2), \end{aligned}$$

where we have used the facts that, for all k, $f_1(x^k) \ge \bar f_1$, $f_2(x^k) \ge \bar f_2$, and $0 < \sigma_{k+1} \le \sigma_k$. Letting $k_1 \to \infty$, we conclude that

$$\sum_{k=0}^{\infty}\delta_{\ell(k)} \le m^{-1}(\sigma_0(f_1(x^0)-\bar f_1) + (f_2(x^0)-\bar f_2)) < +\infty. \tag{3.2}$$

In particular,

$$\delta_{\ell(k)} \to 0 \quad \text{as } k \to \infty. \tag{3.3}$$

Take any $\bar x \in S_1 \ne \emptyset$. Using Lemma 2.1(i), we obtain that

$$\begin{aligned} \|x^{k+1}-\bar x\|^2 &= \|x^k-\bar x\|^2 - \frac{2}{\mu_{\ell(k)}}\langle \hat g^{\ell(k)}, x^k-\bar x\rangle + \frac{1}{\mu_{\ell(k)}^2}\|\hat g^{\ell(k)}\|^2 \\ &\le \|x^k-\bar x\|^2 + \frac{2}{\mu_{\ell(k)}}\left(F_{\sigma_k}(\bar x) - F_{\sigma_k}(x^k) + \hat\varepsilon^{k,\ell(k)} + \frac{1}{2\mu_{\ell(k)}}\|\hat g^{\ell(k)}\|^2\right) \\ &= \|x^k-\bar x\|^2 + \frac{2}{\mu_{\ell(k)}}\delta_{\ell(k)} \\ &\quad + \frac{2}{\mu_{\ell(k)}}\left(\sigma_k(f_1(\bar x)-f_1(x^k)) + f_2(\bar x) - f_2(x^k)\right) \\ &\le \|x^k-\bar x\|^2 + \frac{2}{\mu_{\ell(k)}}\delta_{\ell(k)} + \frac{2\sigma_k}{\mu_{\ell(k)}}\left(f_1(\bar x)-f_1(x^k)\right), \end{aligned} \tag{3.4}$$

where the first inequality is by Lemma 2.1(v), and the last is by the fact that $f_2(\bar x) \le f_2(x^k)$, since $\bar x \in S_1 \subset S_2$.

We next consider separately the following two possible cases:

Case 1. There exists k_2 such that $f_1(\bar x) \le f_1(x^k)$ for all $k \ge k_2$.

Case 2. For each k, there exists $k_3 \ge k$ such that $f_1(\bar x) > f_1(x^{k_3})$.

Case 1. For $k \ge k_2$, we obtain from (3.4) that

$$\|x^{k+1}-\bar x\|^2 \le \|x^k-\bar x\|^2 + \frac{2}{\hat\mu}\delta_{\ell(k)}. \tag{3.5}$$

Recalling (3.2), we conclude that $\{\|x^k-\bar x\|^2\}$ converges (see, e.g., [23, Lem. 2, p. 44]). Hence, $\{x^k\}$ is bounded.

Case 2. For each k, define

$$i_k = \max\{i \le k \mid f_1(\bar x) > f_1(x^i)\}.$$

In the case under consideration, it holds that $i_k \to \infty$ when $k \to \infty$.

We first show that $\{x^{i_k}\}$ is bounded. Observe that

$$\begin{aligned} S_1 &= \{x \in S_2 \mid f_1(x) \le f_1(\bar x)\} \\ &= \{x \in \Re^n \mid \max\{f_2(x)-\bar f_2, f_1(x)-f_1(\bar x)\} \le 0\}. \end{aligned}$$

By assumption, the set S_1 is nonempty and bounded. Therefore, the convex function

$$\phi : \Re^n \to \Re, \quad \phi(x) = \max\{f_2(x) - \bar{f}_2, f_1(x) - f_1(\bar{x})\}$$

has a particular level set $\{x \in \Re^n \mid \phi(x) \leq 0\}$ which is nonempty and bounded. It follows that all level sets of ϕ are bounded (see, e.g., [2, Prop. 2.3.1]), i.e.,

$$L_\phi(c) = \{x \in \Re^n \mid \phi(x) \leq c\}$$

is bounded for any $c \in \Re$.

Since $f_1(x) - \bar{f}_1 \geq 0$ for all $x \in \Re^n$ and $0 < \sigma_{k+1} \leq \sigma_k$, it holds that

$$F_{\sigma_{k+1}}(x) \leq F_{\sigma_k}(x) \qquad \text{for all } x \in \Re^n.$$

Hence,

$$0 \leq F_{\sigma_{k+1}}(x^{k+1}) \leq F_{\sigma_k}(x^{k+1}) \leq F_{\sigma_k}(x^k),$$

where the third inequality follows from (2.10). The above relations show that $\{F_{\sigma_k}(x^k)\}$ is nonincreasing and bounded below. Hence, it converges. It then easily follows that $\{f_2(x^k) - \bar{f}_2\}$ is bounded (because both terms in $F_{\sigma_k}(x^k) = \sigma_k(f_1(x^k) - \bar{f}_1) + (f_2(x^k) - \bar{f}_2)$ are nonnegative).

Fix any $c \geq 0$ such that $f_2(x^k) - \bar{f}_2 \leq c$ for all k. Since $f_1(x^{i_k}) - f_1(\bar{x}) < 0 \leq c$ (by the definition of the index i_k), we have that $x^{i_k} \in L_\phi(c)$, which is a bounded set. This shows that $\{x^{i_k}\}$ is bounded.

By the definition of i_k, it further holds that

$$f_1(\bar{x}) \leq f_1(x^i), \quad i = i_k + 1, \ldots, k \quad (\text{if } k > i_k).$$

Hence, from (3.4), we have that

$$\|x^{i+1} - \bar{x}\|^2 \leq \|x^i - \bar{x}\|^2 + \frac{2}{\hat{\mu}}\delta_{\ell(i)}, \quad i = i_k + 1, \ldots, k.$$

Therefore, for any k, it holds that

$$\begin{aligned} \|x^k - \bar{x}\|^2 &\leq \|x^{i_k} - \bar{x}\|^2 + \frac{2}{\hat{\mu}} \sum_{i=i_k+1}^{k-1} \delta_{\ell(i)} \\ &\leq \|x^{i_k} - \bar{x}\|^2 + \frac{2}{\hat{\mu}} \sum_{i=i_k+1}^{\infty} \delta_{\ell(i)}. \end{aligned} \tag{3.6}$$

Recalling that $i_k \to \infty$, by (3.2) we have that

$$\sum_{i=i_k+1}^{\infty} \delta_{\ell(i)} \to 0 \quad \text{as } k \to \infty. \tag{3.7}$$

Taking also into account the boundedness of $\{x^{i_k}\}$, the relation (3.6) implies that the whole sequence $\{x^k\}$ is bounded.

We next show that all accumulation points of $\{x^k\}$ belong to S_2. For each k, either (2.10) or (2.11) holds. Regardless of whether both conditions hold infinitely often or only one does, it is easy to see that

$$\hat{g}^{\ell(k)} \to 0 \quad \text{and} \quad \hat{\varepsilon}^{k,\ell(k)} \to 0 \quad \text{as } k \to \infty, \tag{3.8}$$

where (3.3) is used if (2.10) holds infinitely often, and (2.11) is used directly.

Let $x \in \Re^n$ be arbitrary but fixed. By Lemma 2.1(v),

$$\sigma_k f_1(x) + f_2(x) \geq \sigma_k f_1(x^k) + f_2(x^k) + \langle \hat{g}^{\ell(k)}, x - x^k \rangle - \hat{\varepsilon}^{k,\ell(k)}. \tag{3.9}$$

Let x^∞ be any accumulation point of $\{x^k\}$. Using boundedness of $\{x^k\}$, the continuity of f_1 and f_2, the fact that $\sigma_k \to 0$ and (3.8), and passing onto the limit in (3.9) along the subsequence which converges to x^∞, we obtain that $f_2(x) \geq f_2(x^\infty)$, where $x \in \Re^n$ is arbitrary. Hence, $x^\infty \in S_2$. □

The rest of the proof is done separately for the following two cases: the number of serious steps when (2.10) is satisfied is either infinite or finite.

THEOREM 3.3. *Let f_1 and f_2 be convex functions such that f_1 is bounded below on $\Re^n$ and the solution set S_1 of problem* (1.1) *is nonempty and bounded.*

Suppose that $\bar{\mu} \geq \mu_\ell \geq \hat{\mu} > 0$ for all iterations ℓ, and that $\mu_{\ell+1} \geq \mu_\ell$ on consecutive null steps.

If serious step descent test (2.10) *is satisfied an infinite number of times and we choose $\{\sigma_k\}$ according to* (1.5) *and $\{\beta_k\} \to 0$ as $k \to \infty$, then $\operatorname{dist}(x^k, S_1) \to 0$ as $k \to \infty$, and all accumulation points of $\{x^k\}$ are solutions of* (1.1).

Proof. Take any $\bar{x} \in S_1$. We again consider separately the two possible cases introduced in the proof of Proposition 3.2:

Case 1. There exists k_2 such that $f_1(\bar{x}) \leq f_1(x^k)$ for all $k \geq k_2$.

Case 2. For each k, there exists $k_3 \geq k$ such that $f_1(\bar{x}) > f_1(x^{k_3})$.

Case 2. Recalling that $i_k = \max\{i \leq k \mid f_1(\bar{x}) > f_1(x^i)\}$ so that $f_1(x^{i_k}) < f_1(\bar{x})$, by the continuity of f_1 it holds that $f_1(x^\infty) \leq f_1(\bar{x})$ for any accumulation point x^∞ of $\{x^{i_k}\}$. Since all accumulation points of $\{x^k\}$ belong to S_2 (as established in Proposition 3.2), it must be the case that all accumulation points of $\{x^{i_k}\}$ are solutions of the problem. In particular,

$$\operatorname{dist}(x^{i_k}, S_1) \to 0 \quad \text{as } k \to \infty. \tag{3.10}$$

For each k, define $\bar{x}^k = P_{S_1}(x^{i_k})$. Using (3.6) with $\bar{x} = \bar{x}^k$ gives

$$\begin{aligned} \operatorname{dist}(x^k, S_1)^2 &\leq \|x^k - \bar{x}^k\|^2 \\ &\leq \operatorname{dist}(x^{i_k}, S_1)^2 + \frac{2}{\hat{\mu}} \sum_{i=i_k+1}^{\infty} \delta_{\ell(i)}. \end{aligned}$$

Passing onto the limit in the latter relation as $k \to \infty$, and using (3.7) and (3.10), we obtain that $\operatorname{dist}(x^k, S_1) \to 0$.

Case 1. As has been shown in Proposition 3.2, in this case the sequence $\{\|x^k - \bar{x}\|\}$ converges for any $\bar{x} \in S_1$. Therefore, if we establish that $\{x^k\}$ has an accumulation point $x^\infty \in S_1$, it would immediately follow that $\{\|x^k - x^\infty\|\} \to 0$; i.e., the whole sequence $\{x^k\}$ converges to $x^\infty \in S_1$.

Suppose first that (2.11) is satisfied only a finite number of times. Suppose further that there is no accumulation point of $\{x^k\}$ which solves (1.1). Since, by Proposition 3.2, all accumulation points are feasible for (1.1), the second assumption means that $\liminf_{k\to\infty} f_1(x^k) > f_1(\bar{x})$, where $\bar{x} \in S_1$. In particular, there exists $t > 0$ such that

$f_1(\bar{x}) \le f_1(x^k) - t$ for all $k \ge k_4$. We then obtain from (3.4) that for $k > k_4$, it holds that

$$\begin{aligned}\|x^{k+1} - \bar{x}\|^2 &\le \|x^k - \bar{x}\|^2 + \frac{2}{\hat{\mu}}\delta_{\ell(k)} - \frac{2t}{\bar{\mu}}\sigma_k \\ &\le \|x^{k_4} - \bar{x}\|^2 + \frac{2}{\hat{\mu}}\sum_{i=k_4-1}^{k}\delta_{\ell(i)} - \frac{2t}{\bar{\mu}}\sum_{i=k_4-1}^{k}\sigma_i.\end{aligned}$$

Passing onto the limit when $k \to \infty$ in the latter relation, we obtain

$$\frac{2t}{\bar{\mu}}\sum_{i=k_4-1}^{\infty}\sigma_i \le \|x^{k_4} - \bar{x}\|^2 + \frac{2}{\hat{\mu}}\sum_{i=k_4-1}^{\infty}\delta_{\ell(i)},$$

which is a contradiction, due to (3.2) and (1.5). Hence, $\liminf_{k\to\infty} f_1(x^k) = f_1(\bar{x})$. Since $\{x^k\}$ is bounded, it must have an accumulation point x^∞ such that $f_1(x^\infty) = f_1(\bar{x})$. As $x^\infty \in S_2$, this means that $x^\infty \in S_1$.

Finally, suppose that (2.11) is satisfied an infinite number of times. Consider the subsequence of indices k for which (2.11) holds (we shall not specify it explicitly) and let $\ell(k)$ denote the associated index ℓ in (2.11). We have that

$$\max\{\sigma_k^{-1}\hat{\varepsilon}^{k,\ell(k)}\,,\,\sigma_k^{-1}\|\hat{g}^{\ell(k)}\|\} \le \beta_k, \quad \beta_k \to 0. \tag{3.11}$$

Taking any $x \in S_2$ and using Lemma 2.1(v), we have that

$$\sigma_k f_1(x) + f_2(x) \ge \sigma_k f_1(x^k) + f_2(x^k) + \langle \hat{g}^{\ell(k)}, x - x^k\rangle - \hat{\varepsilon}^{k,\ell(k)}.$$

Since $f_2(x) \le f_2(x^k)$ for any $x \in S_2$, we obtain

$$f_1(x) \ge f_1(x^k) + \langle \sigma_k^{-1}\hat{g}^{\ell(k)}, x - x^k\rangle - \sigma_k^{-1}\hat{\varepsilon}^{k,\ell(k)}. \tag{3.12}$$

Hence, passing onto the limit in (3.12) as $k \to \infty$ along some subsequence converging to x^∞ and taking into account (3.11), we conclude that $f_1(x) \ge f_1(x^\infty)$ for any $x \in S_2$. Therefore, $x^\infty \in S_1$ also in this case, which concludes the proof. □

It remains to consider the case of a finite number of serious steps in Algorithm 2.1. As already discussed above, this is rather unlikely to occur. Actually, as the next result shows, it can happen only if we hit an exact solution of the problem, which is generally an exceptional situation.

THEOREM 3.4. *Let f_1 and f_2 be convex functions such that f_1 is bounded below on $\Re^n$ and the solution set S_1 of problem (1.1) is nonempty and bounded.*

Suppose that $\bar{\mu} \ge \mu_\ell \ge \hat{\mu} > 0$ for all iterations ℓ, and that $\mu_{\ell+1} \ge \mu_\ell$ on consecutive null steps.

If the serious step descent test (2.10) is satisfied a finite number of times and we choose $\{\sigma_k\} \to 0$ and $\{\beta_k\} \to 0$ as $k \to \infty$, then there exists an iteration index k_0 such that $x^k = x^{k_0}$ for all $k \ge k_0$ and $x^{k_0} \in S_1$.

Proof. Since x^k is changed only when (2.10) holds, it is readily seen that $x^k = x^{k_0}$ for all $k \ge k_0$. By Proposition 3.2, we have that $x^{k_0} \in S_2$.

By Proposition 3.1, we have that for all $k \ge k_0$, σ_k is updated when (2.11) holds. For each k, let $\ell(k)$ denote the index ℓ for which (2.11) is satisfied. We have that

$$\max\{\sigma_k^{-1}\hat{\varepsilon}^{k,\ell(k)}\,,\,\sigma_k^{-1}\|\hat{g}^{\ell(k)}\|\} \le \beta_k, \quad \beta_k \to 0. \tag{3.13}$$

Taking any $x \in S_2$ and using Lemma 2.1(v), we have that

$$\sigma_k f_1(x) + f_2(x) \geq \sigma_k f_1(x^{k_0}) + f_2(x^{k_0}) + \langle \hat{g}^{\ell(k)}, x - x^{k_0} \rangle - \hat{\varepsilon}^{k,\ell(k)}.$$

Since $f_2(x) = f_2(x^{k_0})$ for any $x \in S_2$, we obtain

$$f_1(x) \geq f_1(x^{k_0}) + \langle \sigma_k^{-1} \hat{g}^{\ell(k)}, x - x^{k_0} \rangle - \sigma_k^{-1} \hat{\varepsilon}^{k,\ell(k)}. \tag{3.14}$$

Hence, passing onto the limit in (3.14) as $k \to \infty$ and taking into account (3.13), we conclude that $f_1(x) \geq f_1(x^{k_0})$ for any $x \in S_2$. Therefore, $x^{k_0} \in S_1$, as claimed. ∎

4. Computational experiments. In this section, we report on some numerical experiments for the problem of minimizing a piecewise quadratic convex function over a set defined by monotone linear complementarity constraints. Specifically, we consider the problem

$$\begin{array}{ll} \text{minimize} & \max_{j=1,\ldots,l}\{\langle A^j x, x\rangle + \langle b^j, x\rangle + c^j\} \\ \text{subject to} & Qx + q \geq 0,\ x \geq 0,\ \langle x, Qx + q\rangle \leq 0, \end{array} \tag{4.1}$$

where Q and A^j, $j = 1, \ldots, l$, are $n \times n$ positive semidefinite matrices; q and b^j, $j = 1, \ldots, l$, are vectors in $\Re^n$; and $c^j \in \Re$, $j = 1, \ldots, l$. This problem is converted to the setting of the paper by choosing

$$f_1(x) = \max_{j=1,\ldots,l}\{\langle A^j x, x\rangle + \langle b^j, x\rangle + c^j\},$$

$$f_2(x) = \sum_{i=1}^{n} \max\{-x_i, 0\} + \sum_{i=1}^{n} \max\{-(Qx+q)_i, 0\} + \max\{\langle Qx + q, x\rangle, 0\}.$$

The code is written in MATLAB, essentially by making modifications to a more-or-less standard unconstrained proximal bundle code. Runs are performed under MATLAB Version 7.0.0.19901 (R14). The test problems were constructed by first generating a feasible point $\bar{x}$ of (4.1), and then a function f_1 for which $\bar{x}$ is optimal. Details are presented next.

The process starts with defining an $n \times n$ positive semidefinite matrix Q of rank $r < n$, whose entries are uniformly distributed in the interval $[-5, 5]$. We next generate a point $\bar{x}$, with each coordinate having equal probability of being zero or being uniformly distributed in $[0, 5]$. Finally, we define $q = -Q\bar{x} + \bar{y}$, where a coordinate of $\bar{y}$ is zero if the corresponding coordinate of $\bar{x}$ is positive, while other coordinates of $\bar{y}$ have equal probability of being zero or uniformly generated from $[0, 5]$. As can be easily seen, such $\bar{x}$ is a feasible point for problem (4.1). It does not satisfy strict complementarity and, typically, is not an isolated feasible point (here, it is important that Q is a degenerate matrix). Obviously, $\bar{x}$ is an unconstrained minimizer of the function f_2, i.e., $\bar{x} \in S_2$.

Next, we construct a function f_1 such that $\bar{x}$ is a minimizer of f_1 over S_2. As the constraints in (4.1) do not satisfy a constraint qualification, we can only overestimate the tangent cone $T_{S_2}(\bar{x})$ to S_2 at $\bar{x}$, which gives underestimation of its dual:

$$(T_{S_2}(\bar{x}))^* \supset K = \operatorname{cone}\left(\{-e^i \mid \bar{x}_i = 0\} \cup \{-Q_i \mid \bar{y}_i = 0\} \cup \{q + (Q + Q^\top)\bar{x}\}\right), \tag{4.2}$$

where e^i is the ith element of the canonical basis of $\Re^n$, Q_i is the ith row of the matrix Q, and $\operatorname{cone}(X)$ stands for the conic hull of the set X in $\Re^n$.

We shall construct the needed function f_1 by defining antigradients of pieces of f_1 active at $\bar{x}$ as some elements belonging to the right-hand side of (4.2). This would guarantee the optimality condition

$$0 \in \partial f_1(\bar{x}) + (T_{S_2}(\bar{x}))^*, \tag{4.3}$$

even though the set $(T_{S_2}(\bar{x}))^*$ is not fully known. First, we generate symmetric $n \times n$ positive semidefinite matrices A^j, $j = 1, \ldots, l$, with random entries distributed in $[-5, 5]$. Choosing the number $l_0 \leq l$ of pieces of f_1 active at $\bar{x}$, we next define

$$b^j = -2A^j\bar{x} - u^j, \;\; u^j \in K, \quad j = 1, \ldots, l_0,$$

where elements u^j of K are generated by taking random coefficients in $[0, 1]$ for all vectors in the right-hand side of (4.2). The elements b^j, $j = l_0+1, \ldots, l$, are generated randomly.

It remains to make sure that the first l_0 pieces in the definition of f_1 are active at $\bar{x}$. To this end, we compute

$$\bar{c} = 5 + \max_{j=1,\ldots,l}\{\langle A^j\bar{x}, \bar{x}\rangle + \langle b^j, \bar{x}\rangle\},$$

and set

$$c^j = \bar{c} - \langle A^j\bar{x}, \bar{x}\rangle - \langle b^j, \bar{x}\rangle, \quad j = 1, \ldots, l_0,$$

$$c^j = 0, \quad j = l_0 + 1, \ldots, l.$$

It can be seen that for the point $\bar{x}$, the maximum in the definition of f_1 is attained for indices $j = 1, \ldots, l_0$, and that $f_1(\bar{x}) = \bar{c}$. By the previous constructions, we have that (4.3) holds, and thus $\bar{x}$ is a solution of (4.1). Furthermore, the optimal value of this problem is $\bar{c}$.

Our code is a slightly simplified version of Algorithm 2.1, in particular in the following two details. First, instead of an aggregation technique to control the bundle, we use simple selection of active pieces; i.e., after every iteration we discard those cutting planes which correspond to zero multipliers in the solution of the QP subproblem. Second, we ignore the safeguard (2.11) that detects when the current point x^k is almost a minimizer of F_{σ_k}, and so σ_k needs to be reduced (even if a serious step has not yet been constructed). As already discussed above, since at no iteration F_{σ_k} is being minimized to any specific precision, this situation is unlikely to occur prematurely if σ_k is updated after each serious step. This intuition was confirmed by our experiments. We observed that optimality is achieved only asymptotically, and so the standard bundle stopping test,

$$\hat{\varepsilon}^{k,\ell} \leq t_1 \quad \text{and} \quad \|\hat{g}^\ell\|^2 \leq t_2, \tag{4.4}$$

can be used without any harm. But, of course, one has to be aware that this stopping test cannot be fully reliable in our setting. In our experiments, we set $t_1 = 10^{-2}$ and $t_2 = 10^{-4}$, as it is often difficult to get more precision from a nondifferentiable optimization code in a simple MATLAB implementation. We start with $x^0 = (2, \ldots, 2)$, and set $m = 10^{-1}$ in the descent test (2.10). The proximal parameter μ_ℓ in (2.5) is changed at serious steps only by the safeguarded version of the *reversal quasi-Newton* scalar update; see [3, section 9.3.3]. More precisely,

$$\mu_{k+1} = \min\left\{c_1, \max\{\tilde{\mu}_{k+1}, c_2\}\right\},$$

TABLE 4.1
Summary of numerical experiments.

	Convergence (out of 20)		"Failures" (out of 20)	
$n = 5$	18 cases	38.3 oracle calls	2 cases	100 oracle calls
$\operatorname{rank} Q = 4$	$R_1 = 2.2 * 10^{-5}$	$R_2 = 1.2 * 10^{-5}$	$R_1 = 4.1 * 10^{-4}$	$R_2 = 2.2 * 10^{-4}$
$n = 5$	19 cases	32.2 oracle calls	1 case	100 oracle calls
$\operatorname{rank} Q = 2$	$R_1 = 6.2 * 10^{-4}$	$R_2 = 8.1 * 10^{-5}$	$R_1 = 3.2 * 10^{-4}$	$R_2 = 1.1 * 10^{-5}$
$n = 10$	12 cases	109.5 oracle calls	8 cases	200 oracle calls
$\operatorname{rank} Q = 8$	$R_1 = 2.8 * 10^{-5}$	$R_2 = 1.4 * 10^{-5}$	$R_1 = 2.2 * 10^{-4}$	$R_2 = 3.3 * 10^{-5}$
$n = 10$	14 cases	89.9 oracle calls	6 cases	200 oracle calls
$\operatorname{rank} Q = 5$	$R_1 = 3.7 * 10^{-4}$	$R_2 = 4.2 * 10^{-5}$	$R_1 = 7.2 * 10^{-4}$	$R_2 = 5.3 * 10^{-4}$
$n = 10$	16 cases	60.6 oracle calls	4 cases	200 oracle calls
$\operatorname{rank} Q = 2$	$R_1 = 9.8 * 10^{-4}$	$R_2 = 5.4 * 10^{-6}$	$R_1 = 2 * 10^{-3}$	$R_2 = 3.1 * 10^{-6}$

where $\tilde{\mu}_{k+1}$ is the value prescribed by [3, section 9.3.3], and $c_1 = 10$, $c_2 = 10^{-1}$. Subproblems (2.5) are solved by applying the MATLAB QP routine `qp.m` to the dual formulation of (2.5).

For updating the weight parameter, we use the simple generic choice

$$\sigma_k = \sigma_0/(k+1). \tag{4.5}$$

For lower dimensions (say, $n = 5$), when fewer iterations are expected, we start with $\sigma_0 = 10$. For higher dimensions (say, $n = 10$), when more iterations are typically needed, we start with $\sigma_0 = 20$. We have experimented with other possibilities, like keeping σ_k fixed over some number of serious steps, as well as with some more involved strategies. While improvements are possible, at this time we did not find them significant enough, with respect to the simple (4.5), to warrant their description. Generally, our experiments are intended for merely verifying that the proposed algorithm works and in a reasonable way. We did not spend much time on tuning various parameters to obtain an efficient code. To achieve this, as a first step, one should dispense with the generic `qp.m` MATLAB QP solver, which is known to be problematic (and was observed to be a limitation for our experiments as well). Instead, some good specialized solver (e.g., based on [16, 10]) has to be employed.

Our results are summarized in Table 4.1. We report on problems of dimensions $n = 5$ and $n = 10$, with various degrees of degeneracy of matrix Q, i.e., for different values of rank $Q = r < n$. Note that the number of constraints in (4.1) is $2n + 1$. For each pair of n and r the results are averaged over 20 runs. For all the problems, $l = 5$ and $l_0 = 3$; i.e., f_1 is defined by a maximum of five quadratic functions, with three of them being active at $\bar{x}$. We found that moderate variations of l and l_0 do not change much of the average behavior of the method. We thus keep them fixed in our report, to simplify the table. We report the number of times (out of 20 runs) that convergence had been declared according to the stopping rule (4.4), and the number of times this did not happen (declared as a failure) after a maximum allowed number of calls to the oracle (i.e., evaluations of f_1, f_2, and of their subgradients). In the case of $n = 5$, the maximal number of oracle calls is 100, and in the case of $n = 10$, it is 200. For both outcomes, we report the average number of oracle calls at termination (which is redundant in the case of failures) and the average of the relative accuracies achieved with respect to the optimal value $\bar{c}$ of problem (4.1) and of the (in)feasibility measure (the optimal value of f_1, which is zero). Specifically, in Table 4.1, we denote

$$R_1 = |(f_1(x^k) - \bar{c})/(f_1(x^0) - \bar{c})|, \quad R_2 = f_2(x^k)/f_2(x^0),$$

where x^k is the last serious iterate before termination.

We note that even in the cases of "failure" the method actually makes reasonable progress to the solution of the problem, as evidenced by the values of R_1 and R_2 in Table 4.1. We believe that a more careful implementation, including a better QP solver, should improve the accuracy (especially in higher dimensions) and eliminate "failures" of nonsatisfaction of the stopping rule (4.4). To this end, we observed that in most cases, the values of R_1 and R_2 (which measure actual proximity to solution) are very satisfactory, and close to those reported at termination, well before the stopping rule (4.4) is activated or the maximum number of oracle calls is reached. To some extent, this is quite normal for bundle methods, as they have to generate enough information in order to "recognize" optimality of the current point. For example, even starting with $x^0 = \bar{x}$, about 20 oracle calls were required in our experiments before the method stopped according to (4.4). But in some cases, even when the values in the stopping test (4.4) are already quite close to the required tolerances relatively early, it proves difficult to get more precision and satisfy (4.4). As already stated, we believe that the QP solver used in our implementation is likely the main reason we are not able to progress to higher accuracy with respect to stopping test (4.4). In any case, we believe that Table 4.1 shows reasonable behavior of Algorithm 2.1, even in our simple implementation, on problems with complementarity constraints (which is a difficult class of problems). Finally, we observe that degeneracy of the matrix Q defining complementarity constraints is not a problem for our algorithm at all. Actually, problems with higher degeneracy of Q appear even easier to solve. We conjecture that the reason for this is that, in the case of high degeneracy of Q, the feasible set of (4.1) is larger and the function f_2 is easier to minimize. This may make the overall problem easier to deal with in our setting.

5. Concluding remarks. We have presented a bundle method for solving a nonsmooth convex bilevel problem, which includes standard nonsmooth constrained optimization as a special case. The attractive feature of the method is that it is completely explicit. In particular, it does not require an iterative solution (not even approximate) of any optimization subproblems with general structure. Moreover, in the case of optimization, no constraint qualifications are required for convergence.

Acknowledgment. The author thanks Claudia Sagastizábal for her MATLAB unconstrained bundle code, which served as the basis for the implementation of Algorithm 2.1.

REFERENCES

[1] A. Auslender, *Numerical methods for nondifferentiable convex optimization*, Math. Program. Stud., 30 (1987), pp. 102–126.
[2] D. P. Bertsekas, *Convex Analysis and Optimization*, Athena Scientific, Belmont, MA, 2003.
[3] J. F. Bonnans, J. Ch. Gilbert, C. Lemaréchal, and C. Sagastizábal, *Numerical Optimization: Theoretical and Practical Aspects*, Springer-Verlag, Berlin, 2003.
[4] A. Cabot, *Proximal point algorithm controlled by a slowly vanishing term: Applications to hierarchical minimization*, SIAM J. Optim., 15 (2005), pp. 555–572.
[5] R. Cominetti and M. Courdurier, *Coupling general penalty schemes for convex programming with the steepest descent and the proximal point algorithm*, SIAM J. Optim., 13 (2002), pp. 745–765.
[6] R. Correa and C. Lemaréchal, *Convergence of some algorithms for convex minimization*, Math. Program., 62 (1993), pp. 261–275.
[7] M. Kočvara and J. V. Outrata, *Optimization problems with equilibrium constraints and their numerical solution*, Math. Program., 101 (2004), pp. 119–149.
[8] A. V. Fiacco and G. P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley & Sons, New York, 1968.

[9] R. Fletcher and S. Leyffer, *A Bundle Filter Method for Nonsmooth Nonlinear Optimization*, Numerical Analysis Report NA/195, Department of Mathematcs, The University of Dundee, Scotland, 1999.

[10] A. Frangioni, *Solving semidefinite quadratic optimization problems within nonsmooth optimization problems*, Comput. Oper. Res., 23 (1996), pp. 1099–1118.

[11] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms*, Springer-Verlag, Berlin, 1993.

[12] V. V. Kalashnikov and N. I. Kalashnikova, *Solving two-level variational inequality. Hierarchical and bilevel programming*, J. Global Optim., 8 (1996), pp. 289–294.

[13] E. Karas, A. Ribeiro, C. Sagastizábal, and M. Solodov, *A bundle-filter method for nonsmooth convex constrained optimization*, Math. Program., to appear.

[14] K. C. Kiwiel, *An exact penalty function algorithm for nonsmooth convex constrained minimization problems*, IMA J. Numer. Anal., 5 (1985), pp. 111–119.

[15] K. C. Kiwiel, *Methods of Descent for Nondifferentiable Optimization*, Lecture Notes in Math. 1133, Springer-Verlag, Berlin, 1985.

[16] K. C. Kiwiel, *A method for solving certain quadratic programming problems arising in nonsmooth optimization*, IMA J. Numer. Anal., 6 (1986), pp. 137–152.

[17] K. C. Kiwiel, *A constraint linearization method for nondifferentiable convex minimization*, Numer. Math., 51 (1987), pp. 395–414.

[18] K. C. Kiwiel, *Exact penalty functions in proximal bundle methods for constrained convex nondifferentiable minimization*, Math. Programming, 52 (1991), pp. 285–302.

[19] C. Lemaréchal, A. Nemirovskii, and Yu. Nesterov, *New variants of bundle methods*, Math. Programming, 69 (1995), pp. 111–148.

[20] Z.-Q. Luo, J.-S. Pang, and D. Ralph, *Mathematical Programs with Equilibrium Constraints*, Cambridge University Press, Cambridge, UK, 1996.

[21] R. Mifflin, *An algorithm for constrained optimization with semismooth functions*, Math. Oper. Res., 2 (1977), pp. 191–207.

[22] R. Mifflin, *A modification and extension of Lemarechal's algorithm for nonsmooth minimization*, Math. Programming Stud., 17 (1982), pp. 77–90.

[23] B. T. Polyak, *Introduction to Optimization*, Optimization Software, Inc., Publications Division, New York, 1987.

[24] C. Sagastizábal and M. Solodov, *An infeasible bundle method for nonsmooth convex constrained optimization without a penalty function or a filter*, SIAM J. Optim., 16 (2005), pp. 146–169.

[25] M. V. Solodov, *An explicit descent method for bilevel convex optimization*, J. Convex Anal., 14 (2007), pp. 227–238.

SIAM J. OPTIM.
Vol. 18, No. 1, pp. 260–289

A UNILATERALLY CONSTRAINED QUADRATIC MINIMIZATION WITH ADAPTIVE FINITE ELEMENTS*

KUNIBERT G. SIEBERT† AND ANDREAS VEESER‡

Abstract. We consider obstacle problems where a quadratic functional associated with the Laplacian is minimized in the set of functions above a possibly discontinuous and thin but piecewise affine obstacle. In order to approximate minimum point and value, we propose an adaptive algorithm that relies on minima with respect to admissible linear finite element functions and on an a posteriori estimator for the error in the minimum value. It is proven that the generated sequence of approximate minima converges to the exact one. Furthermore, our numerical results in two and three dimensions indicate that the convergence rate with respect to the number of degrees of freedom is optimal in that it coincides with the one of nonlinear or adaptive approximation.

Key words. unilaterally constrained minimization, obstacle problems, contact problems, thin obstacles, adaptive finite elements, conforming methods, a posteriori error estimates, convergence of adaptive algorithms, optimal convergence rate, full contact

AMS subject classifications. 90C25, 65N12, 65N15, 65N30

DOI. 10.1137/05064597X

1. Introduction and outline.

The solutions of infinite-dimensional minimization problems are in general not computable. In order to get computable approximations, one usually replaces the infinite-dimensional feasible set by a finite-dimensional, or discrete, one. For an important class of minimizations, this can be done with the help of finite elements. Of course, the error of such an approximate minimum is then affected by the choice of the discrete feasible set. There are two ways to decrease the error. In view of the fact that the true minima are unknown, classical or nonadaptive methods enlarge the discrete feasible sets in such a way that any element of the infinite-dimensional feasible set is approximated by elements from discrete feasible sets; see, e.g., [7]. In contrast, adaptive methods aim only at the approximation of solutions and enlarge the discrete feasible sets with the help of information extracted from data and previous approximate minima; see, e.g., [1, 2, 26]. In many examples, in particular in the presence of singularities, adaptive methods lead to a much more efficient use of the computational resources.

The well-known convergence proofs of the classical finite element methods do not apply to the adaptive framework. In the case of the infinite-dimensional quadratic minimizations associated with linear elliptic boundary value problems, there has been recent progress in the analysis of adaptive finite element methods. We mention the convergence results of Dörfler [9] and Morin, Nochetto, and Siebert [15, 16, 17] that are based upon a result which, under certain conditions, guarantees a strict error reduction in one adaptive iteration. Combining this strict error reduction result with a coarsening procedure, Binev, Dahmen, and DeVore [5] ensured that the finite ele-

*Received by the editors November 24, 2005; accepted for publication (in revised form) November 6, 2006; published electronically April 3, 2007. This research was partially supported by Italian Ministry for Education, University and Research (MIUR) Prin 2004 "Metodi numerici avanzati per equazioni alle derivate parziali di interesse applicativo."

http://www.siam.org/journals/siopt/18-1/64597.html

†Institut für Mathematik, Universität Augsburg, Universitätsstraße 14, D-86159 Augsburg, Germany (siebert@math.uni-augsburg.de, http://scicomp.math.uni-augsburg.de/Siebert/).

‡Dipartimento di Matematica, Università degli Studi di Milano, Via C. Saldini 50, I-20133 Milano, Italy (veeser@mat.unimi.it, http://www.mat.unimi.it/users/veeser/).

ment solutions are near-best approximations, while Stevenson [23] established optimal convergence rates in terms of the number of the degrees of freedom (DOFs) without invoking coarsening.

The sole convergence of adaptive finite elements without coarsening has been established also for a nonquadratic but convex minimization by Veeser [25] and an "equality-constrained" quadratic minimization by Bänsch, Morin, and Nochetto [3]. Similarly to Dahlke, Hochmuth, and Urban [8] in the context of wavelets, the latter is based upon an adaptively approximated Uzawa iteration, and thus the iterates are not Galerkin approximations to the original problem.

This article concerns adaptive (Ritz–)Galerkin approximations with finite elements and their convergence for the following infinite-dimensional "inequality-constrained" quadratic minimization. Let $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, be a polyhedral Lipschitz domain and $f \in L_2(\Omega)$ a load term. The lower obstacle is given by a finite sequence of pairs $\{(K_i, \psi_i)\}_{i=1}^n$ such that

- each $K_i \subset \Omega$ is a nondegenerate closed m-simplex, $m \in \{d-1, d\}$,
- their interiors (with respect to the induced topology) are pairwise disjoint,
- each ψ_i is an affine function over K_i satisfying $\psi_i \le 0$ on $\partial\Omega \cap K_i$.

Let u be the typically unknown minimizer of the "inhomogeneous Dirichlet energy"

$$I[v] := \int_\Omega \tfrac{1}{2}|\nabla v|^2 - fv \tag{1.1}$$

in the set

$$\mathcal{F} := \{v \in H_0^1(\Omega) \mid v \ge \psi_i \text{ on } K_i \text{ for } i = 1, \dots, n\}, \tag{1.2}$$

which is nonempty, convex, and closed thanks to the trace theorem. The minimizer u exists, is unique, and is characterized by the variational inequality

$$\forall\, v \in \mathcal{F} \quad \langle \nabla u, \nabla(v-u)\rangle \ge \langle f, v-u\rangle\,, \tag{1.3}$$

where $\langle\cdot,\cdot\rangle$ indicates the L_2-scalar product; see, e.g., [12, Chapter II, Theorem 2.1].

In order to approximate such minimum computionally, in section 2 we design an adaptive algorithm with continuous linear finite elements. The algorithm is based upon an iteration of the following main steps:

$$\text{minimize} \to \text{estimate} \to \text{select} \to \text{include}, \tag{1.4}$$

that is, determine the minimum u_k of I in the current finite element subset $\mathcal{F}_k$ of $\mathcal{F}$ and estimate its error to test if it already meets a prescribed tolerance; if not, select new basis functions that allow for the inclusion of new directions in the next discrete feasible set $\mathcal{F}_{k+1}$.

The steps "estimate" and "select" involve an a posteriori estimator $\mathcal{E}_k$ for the error in the energy minimum

$$I[u_k] - I[u] \ge 0. \tag{1.5}$$

This is a quantity that is, in terms of the approximate minimizer u_k and data, computable and splits into contributions associated with possibly new basis functions. Although some part of the estimator $\mathcal{E}_k$ is related to the heuristical one of Kornhuber [13], it is new and differs from those in [4, 11, 18, 19, 24] in various aspects, e.g., error notion, accumulation of indicators, and range of covered obstacles.

Its derivation in section 3 is based upon the quantity

$$\sup\{\langle -\mathcal{D}_k, \varphi\rangle \mid \varphi \in H^1_0(\Omega) \text{ such that } \|\nabla\varphi\| \le 1,\ u_k + \varphi \in \mathcal{F}\}, \tag{1.6}$$

where $\mathcal{D}_k \in H^{-1}(\Omega) := H^1_0(\Omega)^*$ denotes the derivative of I (or the residual) in the current approximate minimizer u_k,

$$\forall\, \varphi \in H^1_0(\Omega) \quad \langle \mathcal{D}_k, \varphi\rangle = \langle \nabla u_k, \nabla\varphi\rangle - \langle f, \varphi\rangle\,, \tag{1.7}$$

and $\langle\cdot,\cdot\rangle$ stands for the duality pairing of $H^{-1}(\Omega)$ and $H^1_0(\Omega)$ or the L_2-scalar product. Although (1.6) is used in finite-dimensional constrained optimization as criticality measure, it has not yet been used to derive a posteriori error estimators. Further important ingredients of the derivation are the concept of full contact introduced in [11], an adaptation of the projection operators on stars in [17], and the technique for lower bounds from [25].

In section 4 we prove that the sequence of approximate minima $\{u_k\}_k$ produced by the algorithm converges to the true minimum u in the following sense:

$$I[u_k] \to I[u] \qquad \text{and} \qquad u_k \to u \quad \text{in } H^1(\Omega). \tag{1.8}$$

The described algorithm has been implemented within the framework of the finite element toolbox ALBERTA [20, 21]. Our numerical results in section 5 corroborate and complement the theoretical results of sections 3 and 4. In particular, they illustrate properties of the a posteriori error estimator $\mathcal{E}_k$ and address the convergence speed in (1.8) with respect to the number of DOFs, a key ingredient for the overall complexity of the algorithm. In particular for a singular solution, the observed convergence rate is superior to nonadaptive refinement and coincides with the one of nonlinear or adaptive approximation where the exact solution u is supposed to be known, the obstacle is disregarded, and the (Besov) regularity of u essentially determines the convergence rate.

2. Adaptive minimization with linear finite elements. In this section we introduce the adaptive algorithm for the minimization of I in $\mathcal{F}$ with the help of linear finite elements. Notice that, since finite element functions are defined over meshes, the sequence of approximate minima $\{u_k\}_k$ has to be accompanied by a corresponding sequence of meshes $\{\mathcal{T}_k\}_k$. We start with the initialization of iteration (1.4), then present its single steps, and conclude with some elementary properties.

We denote by $L_2(\Omega)$ the set of measurable functions that are square integrable in Ω with respect to the d-dimensional Lebesgue measure $\mathcal{L}^d$, by $H^1_0(\Omega)$ the $L_2(\Omega)$-functions with first weak derivatives in $L_2(\Omega)$ and zero trace on $\partial\Omega$, and by $H^{-1}(\Omega)$ the dual space of $H^1_0(\Omega)$. For any subset $\omega \subset \Omega$ with nonempty interior, we denote the $L_2(\omega)$-norm by $\|\cdot\|_\omega$ and set $\|\cdot\| = \|\cdot\|_\Omega$.

2.1. Initialization. In order to start and execute iteration (1.4), three inputs are needed:

$$\mathsf{tol} \ge 0, \quad \theta \in (0,1), \quad \text{and } \mathcal{T}_0.$$

The first quantity tol is used as tolerance in the stopping test of step "estimate." If $\mathsf{tol} = 0$, then the algorithm stops only when the current approximate minimizer is the exact one.

The second quantity θ is a parameter in step "select" that affects the number of iterations and the way new DOFs are introduced, two ingredients for the complexity of

the algorithm. On one hand, the closer θ to 1, the lower the number of iterations (1.4). On the other hand, the closer θ to 0, the more selective new DOFs will be introduced. In all our numerical experiments in section 5, we have used the compromise $\theta = 0.3$. See also section 2.6 below for an interpretation of this parameter.

The last quantity $\mathcal{T}_0$ is the initial mesh, i.e., a decomposition of Ω into closed triangles (if $d = 2$) or closed tetrahedrons (if $d = 3$) such that

- $\overline{\Omega} = \cup_{T\in\mathcal{T}_0} T$ and
- the intersection $T_1 \cap T_2$ of each pair $T_1, T_2 \in \mathcal{T}_0$ is empty or a common m-subsimplex, $m = 0, \ldots, d$.

Moreover, we suppose that $\mathcal{T}_0$ is subordinated to $\{K_i\}_{i=1}^n$ in the following sense:

$$\text{each } K_i \text{ is a union of (sub)simplices from } \mathcal{T}_0. \tag{2.1}$$

The successive meshes, which are generated in step "include," inherit this property. This property is crucial to guarantee that the discrete feasible sets are nested and subsets of the nondiscrete one $\mathcal{F}$.

Denoting by $\sigma(T)$ the ratio of the minimum diameter of a ball containing T over the maximum diameter of a ball contained in T, the quantity

$$\sigma_0 := \max_{T\in\mathcal{T}_0} \sigma(T) \in [1, \infty) \tag{2.2}$$

is called the shape regularity of $\mathcal{T}_0$. It affects the decisions that are taken in steps "estimate" and "select." In particular, for a fixed tolerance tol, the stopping test may get less stringent if σ_0 gets bigger.

2.2. Minimize with respect to admissible finite element functions. Suppose that the mesh $\mathcal{T}_k$ has been constructed. In order to define the corresponding finite element minimizer u_k, let

$$V_k := \{v \in C(\overline{\Omega}) \mid \forall\, T \in \mathcal{T}_k \ v_{|T} \text{ is affine}\}$$

be the space of linear finite elements over $\mathcal{T}_k$ and define the discrete feasible set in step k as

$$\mathcal{F}_k := \big\{v \in V_k \mid v(z) = 0 \text{ for } z \in \mathcal{N}_k \cap \partial\Omega \text{ and } v(z) \geq \psi^*(z) \text{ for } z \in \mathcal{N}_k\big\}, \tag{2.3}$$

where $\mathcal{N}_k$ indicates the nodes (or vertices) of $\mathcal{T}_k$ and

$$\psi^*(x) := \max\big\{\psi_i(x) \mid i \in \{1, \ldots, n\} \text{ such that } K_i \ni x\big\}, \qquad x \in \overline{\Omega}, \tag{2.4}$$

with the convention $\max \emptyset = -\infty$. Like $\mathcal{F}$, the set $\mathcal{F}_k$ is closed, convex, and nonempty due to $\psi^* \leq 0$ on $\partial\Omega$. Consequently, I has an unique minimizer u_k in $\mathcal{F}_k$ that is characterized by the discrete variational inequality

$$\forall\, v \in \mathcal{F}_k \quad \langle \nabla u_k, \nabla(v - u_k)\rangle \geq \langle f, v - u_k\rangle\,. \tag{2.5}$$

The approximate minimizer u_k can be computed, e.g., by the SOR method with projection analyzed by Elliott [10] or by the multigrid method of Kornhuber [14]. These and other iterative solvers can be conveniently started with the finite element minimizer u_{k-1} of the previous step.

FIG. 1. *Example of a test function φ_S associated with an interior side S (left) and finite element star ω_z with skeleton γ_z (indicated by dashed lines) for $d = 2$ (right).*

2.3. Estimate error in minimum. In order to estimate the error (1.5), we will use hierarchical indicators related to the underlying operator and indicators related to the load term f. We shall omit the iteration counter k for these indicators and similar quantities that are parametrized by an index set depending on k.

We first introduce the hierarchical indicators. Let $\mathcal{S}_k$ denote the set of interior sides (edges or faces in two or three dimensions, respectively) of $\mathcal{T}_k$. With each such side $S \in \mathcal{S}_k$, we associate a test function $\varphi_S \notin V_k$ with the following properties:

- the support of φ_S is contained in $\omega_S := \{T \in \mathcal{T}_k \mid T \supset S\}$;
- φ_S is continuous and piecewise affine over a finer mesh in ω_S that may be generated in step "include" and contains a node x_S in the interior of S (which will be specified in step "include" more precisely);
- φ_S is positive, attains its maximum in x_S, and is normalized: $\|\nabla\varphi_S\| = 1$.

A construction of φ_S is given in the beginning of section 5; see Figure 1 (left) for an example in $d = 2$.

Given φ_S, we define the computable quantities

$$\rho_S := \langle -\mathcal{D}_k, \varphi_S \rangle \qquad \text{and} \qquad d_S := \frac{u_k(x_S) - \psi^*(x_S)}{\varphi_S(x_S)} \geq 0. \tag{2.6}$$

If $\rho_S > 0$, then φ_S is a descent direction, which points upwards, and thus any correction of u_k in the direction of φ_S is admissible. If $\rho_S < 0$, then $-\varphi_S$ is a downwards correcting descent direction. Hence a correction in this direction may be not admissible. The quantity d_S measures the maximal available space for an admissible correction in the direction of $-\varphi_S$. In other words,

$$u_k + \beta\varphi_S \in \mathcal{F} \qquad \Longleftrightarrow \qquad \beta \geq -d_S. \tag{2.7}$$

If we add the test function φ_S to V_k, the corresponding reduction of the approximate minimum value is at least

$$\xi_S := I[u_k] - I[u_k + \alpha_S\varphi_S] = \alpha_S\rho_S - \tfrac{1}{2}\alpha_S^2, \tag{2.8}$$

where α_S is the solution of a one-dimensional constrained quadratic minimization:

$$I[u_k + \alpha_S\varphi_S] = \min_{\beta \geq -d_S} I[u_k + \beta\varphi_S]. \tag{2.9}$$

The *unconstrained* minimization $\min_{\beta\in\mathbb{R}} I[u_k + \beta\varphi_S]$ has the solution $\beta = \rho_S$. Consequently, if $\rho_S \geq -d_S$, then $\alpha_S = \rho_S$ and $\xi_S = \frac{1}{2}\rho_S^2$; otherwise, there holds $\alpha_S = -d_S$ and $\xi_S = d_S|\rho_S| - \frac{1}{2}d_S^2 > \frac{1}{2}d_S|\rho_S|$; see also Figure 2. Note that in the first case ξ_S depends on $|\rho_S|$ in a quadratic manner, while in the second one essentially in a linear

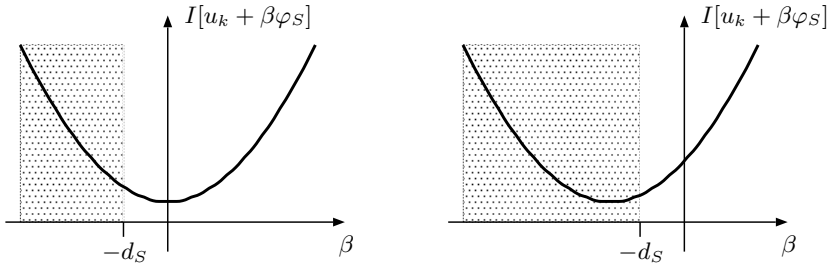


FIG. 2. *The two cases of "unconstrained" (left) and "constrained" (right) new direction.*

manner. The relationship between the reduction ξ_S and the pair (ρ_S, d_S) suggests distinguishing "unconstrained" and "constrained" sides:

$$S \in \mathcal{S}_k^2 \iff \rho_S \geq -d_S \qquad \text{and} \qquad S \in \mathcal{S}_k^1 \iff \rho_S < -d_S \tag{2.10a}$$

such that

$$\xi_S = \tfrac{1}{2}\rho_S^2, \quad S \in \mathcal{S}_k^2, \qquad \text{and} \qquad \xi_S > \tfrac{1}{2} d_S|\rho_S|, \quad S \in \mathcal{S}_k^1. \tag{2.10b}$$

In light of (2.10), we associate ρ_S^2 with an unconstrained side $S \in \mathcal{S}_k^2$ and $d_S|\rho_S|$ with a constrained one $S \in \mathcal{S}_k^1$.

Next, we introduce the notion of full contact and then the indicators related to the load term f. For each node $z \in \mathcal{N}_k$, let $\omega_z := \bigcup\{T \in \mathcal{T}_k \,:\, T \ni z\}$ be the star around z and let $\gamma_z = \bigcup\{S \in \mathcal{S}_k \,:\, S \ni z\}$ be the skeleton of the star ω_z; see Figure 1 (right). Moreover, denote by j_k the normal component of the jumps in ∇u_k across interior sides. More precisely, if $S \in \mathcal{S}_k$, $\omega_S = T^- \cup T^+$ with $T^-, T^+ \in \mathcal{T}_k$, and n is the normal of S pointing from T^- to T^+, then

$$j_{k|S} = \big(\nabla u_{k|T^+} - \nabla u_{k|T^-}\big) \cdot n.$$

The set of full contact nodes is defined as

$$\mathcal{N}_k^0 := \big\{z \in \mathcal{N}_k \mid u_k = \psi_* \text{ in } \omega_z \setminus \partial\omega_z,\ f \leq 0 \text{ in } \omega_z, \text{ and } j_k \leq 0 \text{ on } \gamma_z\big\}, \tag{2.11}$$

where

$$\psi_*(x) := \min\big\{\psi_i(x) \mid i \in \{1, \ldots, n\} \text{ such that } K_i \ni x\big\}, \qquad x \in \overline{\Omega}, \tag{2.12}$$

with the convention $\min \emptyset = +\infty$. Thus u_k is in full contact around a node if it is in contact and additional conditions hold; these additional conditions are necessary for $u_k = u$ around that node. The full contact nodes in turn determine the full contact elements $\mathcal{T}_k^0 := \{T \in \mathcal{T}_k \mid \mathcal{N}_k \cap T \subset \mathcal{N}_k^0\}$ and their complement

$$\mathcal{T}_k^+ := \mathcal{T}_k \setminus \mathcal{T}_k^0. \tag{2.13}$$

Only with "nonfull-contact" elements $T \in \mathcal{T}_k^+$ do we associate indicators $\int_T h_k^2 |f|^2$, where h_k denotes the piecewise constant function with

$$h_{k|T} = \mathcal{L}^d(T)^{1/d}, \qquad T \in \mathcal{T}_k.$$

We finally define the estimator as

$$\mathcal{E}_k := \left(\sum_{S \in \mathcal{S}_k^2} \rho_S^2 \right)^{1/2} + \sum_{S \in \mathcal{S}_k^1} d_S |\rho_S| + \left(\sum_{T \in \mathcal{T}_k^+} \int_T h_k^2 |f|^2 \right)^{1/2}, \tag{2.14}$$

and we stop the algorithm if there holds

$$\mathcal{E}_k \leq \mathsf{tol}. \tag{2.15}$$

The reliability, stability, and efficiency properties of this stopping test are investigated theoretically in sections 3.2 and 3.3, as well as practically in section 5.2 and also in section 5.3.

2.4. Select effective new descent directions. If the stopping test (2.15) is not satisfied, we have to enlarge the discrete feasible set. To this end, we select certain sides for which their corresponding test functions φ_S will be included in V_{k+1}. Denoting the hierarchical part of the estimator by

$$\eta_k := \left(\sum_{S \in \mathcal{S}_k^2} \rho_S^2 \right)^{1/2} + \sum_{S \in \mathcal{S}_k^1} d_S |\rho_S|, \tag{2.16}$$

we choose $\hat{\mathcal{S}}_k \subset \mathcal{S}_k$ such that

$$\left(\sum_{S \in \hat{\mathcal{S}}_k \cap \mathcal{S}_k^2} \rho_S^2 \right)^{1/2} + \sum_{S \in \hat{\mathcal{S}}_k \cap \mathcal{S}_k^1} d_S |\rho_S| \geq \theta \eta_k. \tag{2.17a}$$

Moreover, regarding the third term of the estimator $\mathcal{E}_k$, we define

$$\hat{\mathcal{T}}_k := \left\{ T \in \mathcal{T}_k^+ \mid \int_T h_k^2 |f|^2 \geq \frac{\eta_k^2}{N_k^+} \right\}, \tag{2.17b}$$

where $N_k^+ := \#\mathcal{T}_k^+$ denotes the number of nonfull-contact elements of $\mathcal{T}_k$. In section 4 we will prove that the properties of the estimator (2.14) in combination with the selection criterion (2.17) and the refinement rule (2.18) below imply convergence. Notice that, since $\theta < 1$ in (2.17a), the set $\hat{\mathcal{S}}_k$ of "marked" sides is typically a proper subset of $\mathcal{S}_k$ and not unique. In order to keep the dimension of the next finite element space V_{k+1} low, one should choose $\hat{\mathcal{S}}_k$ such that $\#\hat{\mathcal{S}}_k$ is as small as possible by collecting the biggest indicators. Following an idea of Dörfler [9], such a minimal choice can be approximated with an algorithm that is easy to implement within the data structures of adaptive meshes, has complexity $\#\mathcal{S}_k$, and thus avoids a sort of complexity $\#\mathcal{S}_k \ln \#\mathcal{S}_k$.

2.5. Include selected descent directions. We refine $\mathcal{T}_k$ into $\mathcal{T}_{k+1}$ by bisecting selected elements appropriately. More precisely, given the selected sides $\hat{\mathcal{S}}_k$ and the selected elements $\hat{\mathcal{T}}_k$ from the previous step, we apply the following refinement rule:

(2.18) if an element $T \in \mathcal{T}_k$ belongs to $\hat{\mathcal{T}}_k$ or one of its side to $\hat{\mathcal{S}}_k$, then create by bisection all its children of 2nd (if $d = 2$) or 5th (if $d = 3$) generation.

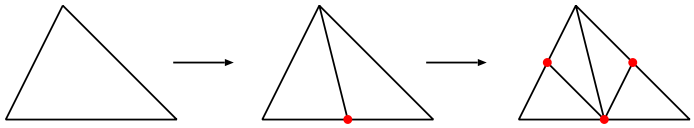


FIG. 3. *A triangle with its children of first and second generation using bisection.*

In order to ensure that the resulting triangulation is conforming, it may happen that additional elements have to be bisected. The required bisections (2.18) can be realized, e.g., with the help of the recursive bisection algorithms described in [21, section 1.1.1]. The refinement by bisection has the following properties:

$$\mathcal{T}_{k+1} \text{ is conforming and its shape regularity is bounded in terms of } \sigma_0, \tag{2.19a}$$

$$\forall\, T \in \mathcal{T}_k \quad h_{k+1|T} < h_{k|T} \implies h_{k+1|T} \le \tfrac{1}{2}\, h_{k|T}, \tag{2.19b}$$

$$V_k \subset V_{k+1}. \tag{2.19c}$$

In two dimensions, the refinement rule (2.18) ensures that for each selected edge $S \in \hat{\mathcal{S}}_k$ the midpoint x_S belongs to the nodes of $\mathcal{T}_{k+1}$; see Figure 3. Now consider a marked face $S \in \hat{\mathcal{S}}_k$ in three dimensions. The refinement rule (2.18) here guarantees that at least the edge which is created first inside this face is also bisected. This generates the vertex $x_S = \frac{1}{4}(a_0 + a_1) + \frac{1}{2}a_2$ as a node of $\mathcal{T}_{k+1}$, where a_0, a_1, a_2 is an appropriate enumeration of the vertices of that face. Hence in two and three dimensions, for each marked side, the associated test function φ_S is in the new finite element space:

$$\forall\, S \in \hat{\mathcal{S}}_k \quad \varphi_S \in V_{k+1}. \tag{2.20}$$

The refinement rule (2.18) enters the convergence proof through the properties (2.19) and (2.20).

Now a new mesh $\mathcal{T}_{k+1}$ has been constructed, and we can increment k and repeat the iteration starting with step "minimize."

2.6. A line search interpretation. We now interpret the described algorithm as a line search method. This interpretation also serves as a guideline for the convergence proof in the following sections.

The algorithm defines a finite or infinite sequence $\{u_k\}_k$ of approximate minimizers, accompanied in particular by sequences of triangulations $\{\mathcal{T}_k\}_k$, of finite element spaces $\{V_k\}_k$, and discrete feasible sets $\{\mathcal{F}_k\}_k$.

In view of (2.1), (2.19c), (2.3), and (2.4), there holds the identity

$$\mathcal{F}_k = \{v \in V_k \mid v = 0 \text{ on } \partial\Omega \text{ and } v \ge \psi_i \text{ on } K_i \text{ for } i = 1, \dots, n\},$$

which in turn yields that the discrete feasible sets are nested and the minimization in each iteration is an inner approximation of (or a conforming method for) the original minimization:

$$\mathcal{F}_k \subset \mathcal{F}_{k+1} \subset \mathcal{F}. \tag{2.21}$$

This implies

$$I[u_k] \ge I[u_{k+1}] \ge I[u]. \tag{2.22}$$

In descent methods, the convergence is usually ensured with the help of a suitable strengthening of the strict decrease $I[u_{k+1}] < I[u_k]$. Here such strict decrease typically holds (see Remark 4.1), but, although there holds $\mathcal{F}_{k+1} \neq \mathcal{F}_k$ whenever $\mathcal{E}_k > 0$, there may be exceptions: for example, if the obstacle is negative and the load term f is $L_2(\Omega)$-orthogonal to the new finite element space V_{k+1} and thus not visible in (2.5), then $I[u_{k+1}] = I[u_k]$; see also [15]. This may happen for any refinement rule that generates a fixed finite number of subelements. We therefore decided to deal with the presence of degenerate iterations with $I[u_{k+1}] = I[u_k]$.

The strengthening of $I[u_{k+1}] < I[u_k]$ in line search methods consists of conditions on the search direction (e.g., the angle condition) and the step length (e.g., Armijo–Goldstein–Wolfe conditions) which imply that the criticality measure for u_k is bounded in terms of the decrease $I[u_{k+1}] - I[u_k]$. Here the derivation of such a bound, as well as the determination of a computable search direction, is complicated by the fact that the derivative $\mathcal{D}_k$ is an infinite-dimensional object and by the presence of the aforementioned degenerate iterations.

In step "estimate," the hierarchical part η_k of the estimator investigates the restriction of $\mathcal{D}_k$ to the enlarged finite element space $\tilde{V}_k := V_k \cup \operatorname{span}\{\varphi_S \mid S \in \mathcal{S}_k\}$, while the nonhierarchical part in particular checks for components of f that are orthogonal to $\tilde{V}_k$; the latter happens only in the nonfull-contact region, which thus plays the role of an active or working set for the nonhierarchical part. The full estimator $\mathcal{E}_k$ bounds the noncomputable criticality measure (1.6); see Proposition 3.2 below.

Steps "select" and "refine" then essentially produce the next finite element space V_{k+1}. The construction is such that the following hold:

- The hierarchical part η_k of the estimator can be shown to be controlled by the energy decrease $I[u_{k+1}] - I[u_k]$; see step 2 in the proof of Theorem 4.1 below. To this end, condition (2.17a), which involves the input parameter θ, replaces the angle condition, with the difference that not $\mathcal{D}_k$ on $H^1_0(\Omega)$ but rather its restriction to $\tilde{V}_k \cap H^1_0(\Omega)$ is taken as a reference.
- The most promising directions of $\tilde{V}_k$ are contained in V_{k+1}.
- The active components of f that are orthogonal to $\tilde{V}_k$ will eventually be resolved.

The fact that u_{k+1} is then the (Ritz–)Galerkin approximation in $\mathcal{F}_{k+1}$ implies that the step length is optimal and the search direction is given by the restriction of $-\mathcal{D}_k$ to $V_{k+1} \cap H^1_0(\Omega)$ and the Lagrange multiplier associated with u_{k+1}. More precisely, the search direction is the Riesz representation of their sum in $V_{k+1} \cap H^1_0(\Omega)$ endowed with the $H^1_0(\Omega)$-scalar product $(v, w) \mapsto \langle \nabla v, \nabla w \rangle$. Furthermore, if the Lagrange multipliers of u_{k+1} and u_k are equal, the search direction corresponding to $u_{k+1} - u_k$ is conjugate with respect to the $H^1_0(\Omega)$-scalar product to all previous search directions.

Thus, iteration (1.4) may be seen as a line search method in which the search direction is adaptively determined with the help of an a posteriori error estimator and the step length is optimal.

3. A posteriori estimator for the error in the minimum. The purpose of this section is twofold: on one hand, we theoretically investigate reliability and efficiency of the stopping test (2.15); on the other hand, we prepare the convergence proof in section 4 by starting to establish the link between the criticality measure (1.6) and the decrease $I[u_k] - I[u_{k+1}]$.

Estimates with "$\preccurlyeq$" instead of "$\leq$" indicate a hidden constant, i.e., $a \preccurlyeq b \Leftrightarrow a \leq Cb$ with a constant $C > 0$, that, if not stated otherwise, depends only on the dimension d and the shape regularity σ_0 of $\mathcal{T}_0$ in (2.2).

3.1. Criticality measure and abstract error control. In the unconstrained minimization $\mathcal{F} = H_0^1(\Omega)$, the a posteriori error analysis is based upon the identities

$$I[u_k] - I[u] = \tfrac{1}{2}\|\nabla(u_k - u)\|^2 = \tfrac{1}{2}\|\mathcal{D}_k\|^2_{H^{-1}(\Omega)}, \tag{3.1}$$

where the dual norm of the residual (1.7),

$$\|\mathcal{D}_k\|_{H^{-1}(\Omega)} := \sup\{\langle \mathcal{D}_k, \varphi\rangle \mid \varphi \in H_0^1(\Omega) \text{ such that } \|\nabla\varphi\| \le 1\}, \tag{3.2}$$

depends only on the approximate minimizer u_k and the load f.

For $\mathcal{F}$ as in (1.2), however, already the first identity in (3.1) may not hold. More specifically, the quadratic nature of I readily leads to

$$I[u_k] - I[u] = \langle \mathcal{D}, u_k - u\rangle + \tfrac{1}{2}\|\nabla(u_k - u)\|^2, \tag{3.3}$$

where $\mathcal{D}$ denotes the derivative of I in the minimizer u. Therefore, (1.3) implies

$$\tfrac{1}{2}\|\nabla(u_k - u)\|^2 \le I[u_k] - I[u], \tag{3.4}$$

while an opposite inequality has to take into account the "linear" term $\langle \mathcal{D}, u_k - u\rangle$ when this term is strictly positive; see also the analog situation illustrated in Figure 2 (right).

In addition, a test function (or correction) φ in (3.2) may correspond to a function that is not admissible; i.e., $u_k + \varphi \in \mathcal{F}_k$ may not hold. We therefore consider

$$\rho_k(-\mathcal{D}_k) := \sup\{\langle -\mathcal{D}_k, \varphi\rangle \mid \varphi \in H_0^1(\Omega) \text{ such that } \|\nabla\varphi\| \le 1,\ u_k + \varphi \in \mathcal{F}\}. \tag{3.5}$$

Notice that ρ_k is a seminorm in $H^{-1}(\Omega)$ that is definite in the set of positive functionals:

- thanks to $u_k \in \mathcal{F}$, it is nonnegative, positively 1-homogeneous, and sublinear (and so convex) but not symmetric;
- if $u_k = \psi_i$ in K_i for some $i \in \{1, \dots, n\}$, there are negative functionals $F \neq 0$ with $\rho_k(F) = 0$;
- there holds

$$\rho_k(F) = 0 \implies \langle F, \varphi\rangle \le 0 \ \forall\ \varphi \in H_0^1(\Omega) \text{ with } \varphi \ge 0. \tag{3.6}$$

With the help of $\rho_k(-\mathcal{D}_k)$, the relationship (3.1) can be replaced as follows.

PROPOSITION 3.1 (abstract error control). *The error in the energy minimum is controlled by $\rho_k(-\mathcal{D}_k)$ from* (3.5)*. More precisely, there holds*

$$\tfrac{1}{2}\min\big\{\rho_k(-\mathcal{D}_k)^2, \rho_k(-\mathcal{D}_k)\big\} \le I[u_k] - I[u] \le \max\big\{\tfrac{1}{2}\rho_k(-\mathcal{D}_k)^2, \rho_k(-\mathcal{D}_k)\big\}.$$

Proof. 1. Similarly to (3.3), we obtain

$$I[u_k] - I[u_k + \alpha\varphi] = \alpha\langle -\mathcal{D}_k, \varphi\rangle - \frac{\alpha^2}{2}\|\nabla\varphi\|^2 \tag{3.7}$$

for $\varphi \in H_0^1(\Omega)$ and $\alpha \in \mathbb{R}$.

2. In order to prove the upper bound for $I[u_k] - I[u]$, we choose $\varphi = u - u_k$ and $\alpha = 1$ in (3.7) to obtain

$$I[u_k] - I[u] = \langle -\mathcal{D}_k, u - u_k\rangle - \frac{1}{2}\|\nabla(u - u_k)\|^2. \tag{3.8}$$

The estimation of $\langle -\mathcal{D}_k, u-u_k\rangle$ in terms of $\rho_k(-\mathcal{D}_k)$ from (3.5) suggests considering two cases.

Case 1: $u_k + \|\nabla(u-u_k)\|^{-1}(u-u_k) \in \mathcal{F}$. Then

$$\langle -\mathcal{D}_k, u-u_k\rangle - \frac{1}{2}\|\nabla(u-u_k)\|^2 \leq \rho_k(\mathcal{D}_k)\|\nabla(u-u_k)\| - \frac{1}{2}\|\nabla(u-u_k)\|^2 \leq \frac{1}{2}\rho_k(-\mathcal{D}_k)^2 \tag{3.9}$$

by using $ab \leq \frac{1}{2}a^2 + \frac{1}{2}b^2$.

Case 2: $u_k + \|\nabla(u-u_k)\|^{-1}(u-u_k) \notin \mathcal{F}$. Then, since $u, u_k \in \mathcal{F}$ implies $u_k + \alpha(u-u_k) \in \mathcal{F}$ for all $\alpha \in [0,1]$, we have $\|\nabla(u-u_k)\| < 1$. Consequently, the definition (3.5) of $\rho_k(-\mathcal{D}_k)$ directly gives

$$\langle -\mathcal{D}_k, u-u_k\rangle - \frac{1}{2}\|\nabla(u-u_k)\|^2 \leq \rho_k(\mathcal{D}_k). \tag{3.10}$$

Inserting the two cases (3.9) and (3.10) into the identity (3.8) then yields the claimed upper bound for $I[u_k] - I[u]$.

3. It remains to show the lower bound for $I[u_k] - I[u]$. To this end, choose $\varphi^* \in H_0^1(\Omega)$ such that

$$\rho_k(-\mathcal{D}_k) = \langle -\mathcal{D}_k, \varphi^*\rangle, \quad \|\nabla\varphi^*\| \leq 1, \quad \text{and} \quad u_k + \varphi^* \in \mathcal{F}.$$

Such φ^* can be constructed with the help of a maximizing sequence, weak compactness, and Mazur's lemma. We again consider two cases.

Case 1: $u_k + \rho_k(-\mathcal{D}_k)\varphi^* \in \mathcal{F}$. This, (3.7) with $\varphi = \varphi^*$, and $\alpha = \rho_k(-\mathcal{D}_k)$, as well as the properties of φ^*, imply

$$I[u_k] - I[u] \geq I[u_k] - I[u_k + \rho_k(-\mathcal{D}_k)\varphi^*] = \rho_k(-\mathcal{D}_k)^2 - \frac{\rho_k(-\mathcal{D}_k)^2}{2}\|\nabla\varphi^*\|^2 \geq \frac{1}{2}\rho_k(-\mathcal{D}_k)^2. \tag{3.11}$$

Case 2: $u_k + \rho_k(-\mathcal{D}_k)\varphi^* \notin \mathcal{F}$. Then, in view of $u_k + \varphi^*, u_k \in \mathcal{F}$, we derive $\rho_k(-\mathcal{D}_k) > 1$, and so

$$I[u_k] - I[u] \geq I[u_k] - I[u_k + \varphi^*] = \rho_k(-\mathcal{D}_k) - \frac{1}{2}\|\nabla\varphi^*\|^2 \geq \frac{1}{2}\rho_k(-\mathcal{D}_k) \tag{3.12}$$

by using also (3.7) with $\varphi = \varphi^*$, $\alpha = 1$, and the properties of φ^*.

Combining the inequalities (3.11) and (3.12) yields the claimed lower bound for $I[u_k] - I[u]$. ∎

Suppose for a moment that $I[v] = \frac{1}{2}|v|^2$, $v \in \mathbb{R}$, and that $u_k \in \mathbb{R}$ is *some* approximation of the exact minimum $u = 0$. In this case, straightforward calculations for different cases in the spirit of Figure 2 show that the bounds in the corresponding one-dimensional counterpart of Proposition 3.1 are sharp. Consequently, a proof of improved bounds has to rely on additional information about u_k. The following remark, which will be useful in section 5.4, is an example for such an improvement.

Remark 3.1 (more stringent error control). Suppose that $u_k = \psi_i$ on K_i for all $i = 1, \ldots, n$ and that $\rho_k(-\mathcal{D}_k) \leq 1$. After the computation of u_k, the first condition can be directly checked and the second one may be verified with the help of Proposition 3.2 below. The first condition implies that $u - u_k \geq 0$ in $\bigcup_{i=1}^n K_i$. Consequently, Case 1 in

the proof of the upper bound in Proposition 3.1 applies, and there holds $I[u_k]-I[u] \leq \frac{1}{2}\rho_h(-\mathcal{D}_k)^2$. The second condition excludes Case 2 of the proof of the lower bound since Case 2 entails $\rho_k(-\mathcal{D}_k) > 1$. Thus, there holds also $I[u_k] - I[u] \geq \frac{1}{2}\rho_h(-\mathcal{D}_k)^2$ by the corresponding Case 1. To summarize, under the aforementioned assumptions, we perfectly mimic (3.1) by

$$I[u_k] - I[u] = \tfrac{1}{2}\rho_k(-\mathcal{D}_k)^2. \tag{3.13}$$

3.2. On reliability. The goal of this section is to derive an upper bound of the error $I[u_k]-I[u]$ in the energy minimum in terms of the a posteriori estimator $\mathcal{E}_k$ from (2.14). To this end, in view of Proposition 3.1, we may estimate $\rho_k(-\mathcal{D}_k)$ by $\mathcal{E}_k$, that is, by "local" and, partially, "finite-dimensional" quantities, while the evaluation of $\rho_k(-\mathcal{D}_k)$ corresponds to an infinite-dimensional optimization that involves, in addition to the usual $H^1_0(\Omega)$-constraint, also one of pointwise nature.

A device for "localization" is the partition of unity

$$\sum_{z\in\mathcal{N}_k} \phi_z = 1 \qquad \text{in } \Omega, \tag{3.14}$$

where ϕ_z, $z \in \mathcal{N}_k$, are the hat functions defined by $\phi_z \in V_k$ and $\phi_z(y) = \delta_{yz}$ for all $y \in \mathcal{N}_k$. Notice that the support of ϕ_z is the star around $z \in \mathcal{N}_k$, i.e., $\operatorname{supp}\phi_z = \omega_z$.

Additionally, we need a projection operator Π which allows us, together with the fact that u_k is piecewise affine, to reduce the high flexibility of a generic test function in the definition of $\rho_k(-\mathcal{D}_k)$. The operator Π has to have local properties that are subordinated to the localization induced by the partition of unity (3.14). We define $\Pi : H^1_0(\Omega) \to \operatorname{span}\{\varphi_S \mid S \in \mathcal{S}_k\}$ by requiring, for all $\varphi \in H^1_0(\Omega)$,

$$\forall\, S \in \mathcal{S}_k \quad \int_S \varphi = \int_S \Pi\varphi. \tag{3.15}$$

The latter is, writing $\Pi\varphi = \sum_{S\in\mathcal{S}_k} \alpha_S(\varphi)\varphi_S$, equivalent to

$$\forall\, S \in \mathcal{S}_k \quad \alpha_S(\varphi) = \frac{\int_S \varphi}{\int_S \varphi_S}.$$

To state the local properties of Π, we introduce the following notion: given a side $S \in \mathcal{S}_k$, a function $\varphi \in H^1_0(\Omega)$ is *S-admissible* whenever

$$\forall\, i \in \{1,\dots,n\} \text{ with } S \subset K_i, \text{ there holds } \int_S (u_k + \varphi) \geq \int_S \psi_i. \tag{3.16}$$

Note that $u_k + \varphi \in \mathcal{F}$ implies that φ is S-admissible for all $S \in \mathcal{S}_k$.

LEMMA 3.1 (local properties of Π). *The coefficients α_S, $S \in \mathcal{S}_k$, of Π are stable and monotone on $H^1_0(\Omega)$. More precisely, for all $S \in \mathcal{S}_k$ there holds*

$$\begin{aligned} |\alpha_S(\varphi)| &\preccurlyeq \operatorname{diam}(\omega_S)^{-1}\|\varphi\|_{\omega_S} + \|\nabla\varphi\|_{\omega_S}, \\ \varphi \geq 0 \text{ on } S &\implies \alpha_S(\varphi) \geq 0, \\ \varphi \text{ is } S\text{-admissible} &\implies \alpha_S(\varphi) \succcurlyeq -d_S. \end{aligned}$$

Proof. 1. To show the first claim, we set $h_S := \operatorname{diam}(\omega_S)$ and observe that, thanks to (2.19a), the measures of elements and sides involved in the definition of φ_S are about h_S^d and h_S^{d-1}, respectively. The piecewise affine function φ_S attains

its maximum in x_S and equals 0 outside ω_S. The normalization $\|\nabla\varphi_S\|_{\omega_S} = 1$ now readily implies $h_S^{1-d/2} \preccurlyeq \varphi_S(x_S) \preccurlyeq h_S^{1-d/2}$. From this, one directly computes the lower bound

$$\int_S \varphi_S \succcurlyeq h_S^{d/2}. \tag{3.17}$$

Moreover, if $T \in \mathcal{T}_k$ with $T \supset S$, there holds the "scaled" trace theorem

$$\|\varphi\|_S \preccurlyeq h_S^{-1/2}\|\varphi\|_T + h_S^{1/2}\|\nabla\varphi\|_T. \tag{3.18}$$

The scaling of the norms can be determined by transforming to a reference element, applying there the corresponding trace inequality, and transforming back; see also [7, section 15]. Combining (3.18) with the Cauchy–Schwarz inequality on S yields

$$\int_S \varphi \le h_S^{(d-1)/2}\|\varphi\|_S \preccurlyeq h_S^{d/2}\big(h_S^{-1}\|\varphi\|_{\omega_S} + \|\nabla\varphi\|_{\omega_S}\big),$$

and therefore, in view of (3.17), the first claim is proven.

2. The second claim readily follows from (3.17) and the monotonicity of the integral $\int_S$.

3. It remains to prove the last claim. Suppose that φ is S-admissible. If there is no $i \in \{1,\dots,n\}$ with $K_i \supset S$, then $d_S = \infty$ and the claim is trivial. Otherwise, fix any $i \in \{1,\dots,n\}$ with $K_i \supset S$ and observe

$$\alpha_S(\varphi) \ge \frac{\int_S(\psi_i - u_k)}{\int_S \varphi_S} = -\frac{\int_S(u_k - \psi_i)}{\int_S \varphi_S}.$$

Thus, if we can show

$$\int_S (u_k - \psi_i) \preccurlyeq \big[u_k(x_S) - \psi_i(x_S)\big]\,\mathcal{H}^{d-1}(S), \tag{3.19}$$

then the third claim follows because of $\mathcal{H}^{d-1}(S) \le h_S^{d-1}$, $\varphi_S(x_S) \preccurlyeq h_S^{1-d/2}$, (3.17), and the definition (2.4) of ψ^*.

For $d = 2$, the node x_S is the midpoint of S, and thus (3.19) is a consequence of the midpoint rule since $u_k - \psi_i$ is affine on S. For $d = 3$, the verification of (3.19) is a bit more involved. Denoting by a_0, a_1, a_2 an appropriate enumeration of the vertices of S, we can write $x_S = \frac{1}{4}(a_0 + a_1) + \frac{1}{2}a_2$; see section 2.5. A calculation yields that x_S is the barycenter of S with respect to the measure induced by ϕ_{a_2}, i.e., $x_S = \int_S x\phi_{a_2}(x)dx$, where $\phi_{a_2} \in V_k$ is the hat function at node a_2. Consequently,

$$g \text{ affine } \implies \int_S g\phi_{a_2} = g(x_S)\int_S \phi_{a_2} = g(x_S)\frac{\mathcal{H}^{d-1}(S)}{d}.$$

We now observe $u_k - \psi_i \ge 0$ on S because of $u_k \in \mathcal{F}$, and so the equivalence of norms on the affine functions over a reference side yields

$$\int_S u_k - \psi_i \preccurlyeq \int_S [u_k - \psi_i]\phi_{a_2} = [u_k(x_S) - \psi_i(x_S)]\frac{\mathcal{H}^{d-1}(S)}{d},$$

which is (3.19). Thus, the proof of the last claim is also concluded. ∎

For given $\varphi \in H^1_0(\Omega)$ and $z \in \mathcal{N}_k$, we will apply the projection operator Π to the localized function $\varphi\,\phi_z$:

$$\Pi(\varphi\,\phi_z) = \sum_{S \in \mathcal{S}_z} \alpha_S(\varphi\,\phi_z)$$

with $\mathcal{S}_z := \{S \in \mathcal{S}_k \mid S \ni z\}$ being the set of interior sides in the star ω_z. The first claim of Lemma 3.1, combined with $\|\varphi_S\|_{\omega_S} \preccurlyeq \operatorname{diam}(\omega_S)$, ensures

$$\begin{aligned} \|\Pi(\varphi\,\phi_z)\|_{\omega_z} &\preccurlyeq \|\varphi\,\phi_z\|_{\omega_z} + \operatorname{diam}(\omega_z)\|\nabla(\varphi\,\phi_z)\|_{\omega_z} \\ &\preccurlyeq \|\varphi\|_{\omega_z} + \operatorname{diam}(\omega_z)\|\nabla\varphi\|_{\omega_z}. \end{aligned} \tag{3.20}$$

The last inequality follows from $\|\nabla(\varphi\,\phi_z)\|_{\omega_z} \preccurlyeq \operatorname{diam}(\omega_z)^{-1}\|\varphi\|_{\omega_z} + \|\nabla\varphi\|_{\omega_z}$ since $\phi_z \le 1$ and $|\nabla\phi_z| \preccurlyeq \operatorname{diam}(\omega_z)^{-1}$.

The first bound in Lemma 3.1 and also (3.20) involve not only an H^1-seminorm but also an L_2-norm. In order to estimate the latter ones by the previous ones locally, we shall invoke a suitable Poincaré inequality. It is based upon a cancellation on one of the (not necessarily interior) sides of $\mathcal{T}_k$ containing z. The set of all these sides is indicated with $\bar{\mathcal{S}}_z$. Notice that $\int_S \phi_z = \frac{1}{d}\,\mathcal{H}^{d-1}(S) > 0$ holds for all $S \in \bar{\mathcal{S}}_z$.

LEMMA 3.2 (Poincaré inequality). *Let ω_z be any star of $\mathcal{T}_k$ and $\varphi \in H^1(\omega_z)$. If $\int_S \varphi\phi_z = 0$ for some $S \in \bar{\mathcal{S}}_z$, then there holds*

$$\|\varphi\|_{\omega_z} \preccurlyeq \operatorname{diam}(\omega_z)\|\nabla\varphi\|_{\omega_z}.$$

Proof. Let $S \in \bar{\mathcal{S}}_z$ and suppose $\int_S \varphi\phi_z = 0$. Then we may write

$$\varphi = \varphi - \left[\int_S \phi_z\right]^{-1} \int_S \varphi\phi_z = (\varphi - c) - \left[\int_S \phi_z\right]^{-1} \int_S (\varphi - c)\phi_z,$$

where c is any real. We thus obtain

$$\begin{aligned} \|\varphi\|_{\omega_z} &\preccurlyeq \|\varphi - c\|_{\omega_z} + \frac{\mathcal{L}^d(\omega_z)^{1/2}}{\left[\int_S \phi_z\right]^{1/2}} \left(h_S^{-1/2}\|\varphi - c\|_{\omega_S} + h_S^{1/2}\|\nabla\varphi\|_{\omega_S}\right) \\ &\preccurlyeq \|\varphi - c\|_{\omega_z} + \operatorname{diam}(\omega_z)\|\nabla\varphi\|_{\omega_z} \end{aligned}$$

with the help of a "weighted" Cauchy–Schwarz inequality on S, $\phi_z \le 1$, the scaled trace theorem (3.18), $\mathcal{L}^d(\omega_z) \preccurlyeq \operatorname{diam}(\omega_z)^d$, $\int_S \phi_z \succcurlyeq h_S^{d-1}$, and $h_S \approx \operatorname{diam}(\omega_z)$. The variant [22, eq. (4.2)] of the Bramble–Hilbert lemma then finishes the proof. ∎

After these preparations we now turn to the upper bound for $\rho_h(-\mathcal{D}_k)$. We refer to a node $z \in \mathcal{N}_k \setminus \partial\Omega$ as a *proper contact node* if $u_k(z) = \psi^*(z)$ and (2.5) is "locally" strict in that $\langle \nabla u_k, \nabla\phi_z\rangle > \langle f, \phi_z\rangle$ holds. If in addition there is no side $S \in \mathcal{S}_z$ and no $i \in \{1, \dots, n\}$ such that $S \subset K_i$ and $u_k = \psi_i$ on S, we call z an *isolated* proper contact node and set

$$\mathcal{N}_k^{\perp} := \{z \in \mathcal{N}_k \setminus \partial\Omega \mid z \text{ is an isolated proper contact node}\}. \tag{3.21}$$

PROPOSITION 3.2 (upper bound for criticality measure). *We have*

$$\rho_k(-\mathcal{D}_k) \preccurlyeq \mathcal{E}_k,$$

where the hidden constant depends on d and σ_0 and, only if $\mathcal{N}_k^{\perp} \neq \emptyset$, in addition on f, $\{\psi_i\}_{i=1}^n$, and $\mathcal{T}_0$.

Proof. 1. Let $\varphi \in H^1_0(\Omega)$ such that $\|\nabla\varphi\| \le 1$ and $u_k + \varphi \in \mathcal{F}$. We have to show that $\langle -\mathcal{D}_k, \varphi\rangle \preccurlyeq \mathcal{E}_k$. To this end, we derive the representation formula

$$\begin{aligned}\langle -\mathcal{D}_k, \varphi\rangle &= \sum_{T\in\mathcal{T}_k} \int_T \big(f\varphi - \nabla u_k\cdot\nabla\varphi\big) = \sum_{S\in\mathcal{S}_k}\int_S j_k\varphi + \sum_{T\in\mathcal{T}_k}\int_T f\varphi \\ &= \sum_{z\in\mathcal{N}_k}\left(\int_{\gamma_z} j_k\varphi\phi_z + \int_{\omega_z} f\varphi\phi_z\right)\end{aligned} \tag{3.22}$$

with the help of elementwise integration by parts and the partition of unity (3.14). We set

$$\rho_z(\varphi) := \int_{\gamma_z} j_k\varphi\phi_z + \int_{\omega_z} f\varphi\phi_z, \qquad z\in\mathcal{N}_k, \tag{3.23}$$

and observe that the definition of u_k implies, for any interior node $z \in \mathcal{N}_k\cap\Omega$,

$$\rho_z(1) \le 0, \tag{3.24a}$$

$$u_k(z) > \psi^*(z) \implies \rho_z(1) = 0. \tag{3.24b}$$

In fact, (3.24a) follows by choosing $v = u_k + \phi_z$ in (2.5), while (3.24b) by choosing $v = u_k - \epsilon\phi_z$ with $\epsilon > 0$ sufficiently small. In what follows, we shall estimate the terms $\rho_z(\varphi)$, $z\in\mathcal{N}_k$, separately, depending on the node type.

2. We consider the full contact nodes (2.11). If $z\in\mathcal{N}_k^0$, then $\omega_z = \bigcup_{i\in I_z} K_i\cap\omega_z$ with $I_z := \{i\in\{1,\dots,n\} \mid K_i\cap\omega_z \ne \emptyset\}$; for otherwise there exists a $x\in\omega_z\setminus\partial\omega_z$ such that $\psi_*(x) = \infty$, which is in contradiction with $\psi_*(x) = u_k(x) < \infty$. In view of $u_k + \varphi \in \mathcal{F}$, we also have

$$\varphi \ge \psi_i - u_k \ge \psi_* - u_k = 0$$

on $K_i\cap\omega_z$ for any $i\in I_z$ and, thus, $\varphi\ge 0$ on the whole star ω_z. Consequently, the sign conditions on j_k and f in (2.11) yield

$$\rho_z(\varphi) = \int_{\gamma_z} j_k\varphi\phi_z + \int_{\omega_z} f\varphi\phi_z \le 0 \qquad \text{for any } z\in\mathcal{N}_k^0. \tag{3.25}$$

Because $\rho_z(\varphi)$ is nonpositive, these values do not contribute to the upper bound.

3. For the remaining nodes $z\in\mathcal{N}_k\setminus\mathcal{N}_k^0$, we first derive the following conditional auxiliary estimate for $\zeta\in H^1(\omega_z)$:

If $\zeta\phi_z$ is S-admissible $\forall\ S\in\mathcal{S}_z$, then $\rho_z(\zeta)$ is bounded by

$$\sum_{S\in\mathcal{S}_z\cap\mathcal{S}_k^1} d_S|\rho_S| + \left(\sum_{S\in\mathcal{S}_z\cap\mathcal{S}_k^2} |\rho_S|^2 + \|h_k f\|^2_{\omega_z}\right)^{1/2}\left(\frac{\|\zeta\|_{\omega_z}}{\operatorname{diam}(\omega_z)} + \|\nabla\zeta\|_{\omega_z}\right). \tag{3.26}$$

Exploiting (3.15) and the fact that j_k is constant on any side, we may write

$$\begin{aligned}\rho_z(\zeta) &= \int_{\gamma_z} j_k\zeta\phi_z + \int_{\omega_z} f\zeta\phi_z \\ &= \int_{\gamma_z} j_k\Pi(\zeta\phi_z) + \int_{\omega_z} f\Pi(\zeta\phi_z) + \int_{\omega_z} f\big[\zeta\phi_z - \Pi(\zeta\phi_z)\big]\end{aligned}$$

such that we obtain

$$\begin{aligned}\int_{\gamma_z} j_k\Pi(\zeta\phi_z) + \int_{\omega_z} f\Pi(\zeta\phi_z) &= \langle -\mathcal{D}_k, \Pi(\zeta\phi_z)\rangle = \sum_{S\in\mathcal{S}_z} \alpha_S(\zeta\phi_z)\rho_S \\ &= \sum_{S\in\mathcal{S}_z\cap\mathcal{S}_k^1} \big[-\alpha_S(\zeta\phi_z)\big]|\rho_S| + \sum_{S\in\mathcal{S}_z\cap\mathcal{S}_k^2} \alpha_S(\zeta\phi_z)\rho_S \\ &\preccurlyeq \sum_{S\in\mathcal{S}_z\cap\mathcal{S}_k^1} d_S|\rho_S| + \sum_{S\in\mathcal{S}_z\cap\mathcal{S}_k^2} |\rho_S|\big[\operatorname{diam}(\omega_S)^{-1}\|\zeta\|_{\omega_S} + \|\nabla\zeta\|_{\omega_S}\big]\end{aligned}$$

by Lemma 3.1 and $\|\nabla(\zeta\phi_z)\|_{\omega_S} \preccurlyeq \operatorname{diam}(\omega_S)^{-1}\|\zeta\|_{\omega_S} + \|\nabla\zeta\|_{\omega_S}$. In addition,

$$\int_{\omega_z} f\big[\zeta\phi_z - \Pi(\zeta\phi_z)\big] \preccurlyeq \|h_k f\|_{\omega_z}\big[\operatorname{diam}(\omega_z)^{-1}\|\zeta\|_{\omega_z} + \|\nabla\zeta\|_{\omega_z}\big]$$

by (3.20) and $h_k \succcurlyeq \operatorname{diam}(\omega_z)$ on ω_z. Summing up the two estimates gives the claimed bound (3.26) for $\rho_z(\zeta)$.

4. In order to apply (3.26) and then Lemma 3.2, it is convenient to write

$$\rho_z(\varphi) = \rho_z(\varphi^+) + \rho_z(\varphi^-), \tag{3.27}$$

where $\varphi^+ = \max\{\varphi, 0\}$ and $\varphi^- = \min\{\varphi, 0\}$ denote the positive and negative part of φ, respectively. Since φ^+ is the "unconstrained" part of the test function φ, the corresponding terms $\rho_z(\varphi^+)$ should be treated in a similar way as in an unconstrained problem. We first estimate these terms and then the ones associated with φ^-. Let $z \in \mathcal{N}_k \setminus \mathcal{N}_k^0$ be arbitrary. Then there exist a nonnegative real $c_z^+ \geq 0$ and a side $S_z \in \bar{\mathcal{S}}_z$ in ω_z such that

$$\int_{S_z} (\varphi^+ - c_z^+)\phi_z = 0. \tag{3.28}$$

In fact, if $z \in \partial\Omega$ is a boundary vertex, then take $c_z^+ = 0$ and a side in $\partial\omega_z \cap \partial\Omega$ for S_z; otherwise, choose

$$c_z^+ := \min\left\{ \left[\int_S \phi_z\right]^{-1} \int_S \varphi^+\phi_z \mid S \in \mathcal{S}_z \right\} \geq 0$$

together with the minimizing side S_z. By construction of c_z^+, the locally shifted test function $\zeta\phi_z$ with $\zeta = \varphi^+ - c_z^+$ is S-admissible for all $S \in \mathcal{S}_z$. Exploiting (3.24a), applying (3.26) and Lemma 3.2 with (3.28), we derive

$$\begin{aligned}\rho_z(\varphi^+) &= \rho_z(\varphi^+ - c_z^+) + c_z^+\rho_z(1) \leq \rho_z(\varphi^+ - c_z^+) \\ &\preccurlyeq \sum_{S\in\mathcal{S}_z\cap\mathcal{S}_k^1} d_S|\rho_S| + \left(\sum_{S\in\mathcal{S}_z\cap\mathcal{S}_k^2} |\rho_S|^2 + \|h_k f\|_{\omega_z}^2\right)^{1/2} \|\nabla\varphi^+\|_{\omega_z}.\end{aligned} \tag{3.29}$$

Notice that (3.28) and the sign $c_z^+\rho_z(1) \leq 0$ are crucial for deriving (3.29).

5. Next, we estimate those $\rho_z(\varphi^-)$, $z \in \mathcal{N}_k \setminus \mathcal{N}_k^0$, for which we can proceed similarly as for $\rho_z(\varphi^+)$. More precisely, we suppose that there exist a nonpositive real $c_z^- \leq 0$ and a side $S_z \in \bar{\mathcal{S}}_z$ in ω_z such that

$$\int_{S_z} (\varphi^- - c_z^-)\phi_z = 0 \quad \text{and} \quad c_z^-\,\rho_z(1) \leq 0. \tag{3.30}$$

Then (3.26) with $\zeta = \varphi^- - c_z^-$ and Lemma 3.2 yield

$$\rho_z(\varphi^-) = \rho_z(\varphi^- - c_z^-) + c_z^-\rho_z(1) \le \rho_z(\varphi^- - c_z^-) \tag{3.31}$$
$$\preccurlyeq \sum_{S\in\mathcal{S}_z\cap\mathcal{S}_k^1} d_S|\rho_S| + \left(\sum_{S\in\mathcal{S}_z\cap\mathcal{S}_k^2} |\rho_S|^2 + \|h_k f\|_{\omega_z}^2\right)^{1/2} \|\nabla\varphi^-\|_{\omega_z}.$$

Condition (3.30) is verified for boundary nodes with $c_z^- = 0$ and a boundary side S_z. For interior nodes z with $\rho_z(1) = 0$, we can choose, for instance,

$$c_z^- := \max\left\{\left[\int_S \phi_z\right]^{-1} \int_S \varphi^-\phi_z \mid S\in\mathcal{S}_z\right\} \le 0$$

and S_z as the maximizing side.

In view of (3.24), the remaining nodes are necessarily proper contact nodes. For contact nodes that are not isolated, there exist a side $S_z \in \mathcal{S}_z$ and some $i \in \{1,\dots,n\}$ such that

$$0 \ge \varphi^- \ge \psi_i - u_k = 0 \quad \text{on } S_z$$

and, thus, (3.30) holds with $c_z^- = 0$.

6. It remains to estimate $\rho_z(\varphi^-)$ for isolated proper contact nodes $z \in \mathcal{N}_k^\perp$. In order to compensate a missing cancellation as in (3.30), we observe that for isolated proper contact nodes there is some $i \in \{1,\dots,n\}$ and some side $S \in \mathcal{S}_z$ such that

$$u_k(z) = \psi_i(z) \quad \text{and} \quad u_k + \varphi^- \ge \psi_i \text{ on } S$$

because z is in contact. Hence, by a "discrete" Poincaré inequality and $h_S\,\mathcal{H}^{d-1}(S) \preccurlyeq \mathcal{L}^d(\omega_S)$,

$$\|\varphi^-\|_S \le \|\psi_i - u_k\|_S \preccurlyeq h_S\|\nabla_S(\psi_i - u_k)\|_S$$
$$\preccurlyeq \operatorname{diam}(\omega_S)^{1/2}\Big(\|\nabla u_k\|_{\omega_S} + \mathcal{L}^d(\omega_S)^{1/2}M\Big),$$

where ∇_S is the tangential gradient on S and $M := \max\{|\nabla\psi_i| \mid i = 1,\dots,n\}$. Therefore $c_z^- := \big[\int_S \phi_z\big]^{-1}\int_S \varphi^-\phi_z$ satisfies

$$\|c_z^-\|_{\omega_z} \le \frac{\mathcal{L}^d(\omega_z)^{1/2}}{\big[\int_S \phi_z\big]^{1/2}}\|\varphi^-\|_S \preccurlyeq \operatorname{diam}(\omega_z)\big[\|\nabla u_k\|_{\omega_z} + \mathcal{L}^d(\omega_z)^{1/2}M\big],$$

whence

$$\|\varphi^-\|_{\omega_z} \le \|\varphi^- - c_z^-\|_{\omega_z} + \|c_z^-\|_{\omega_z}$$
$$\preccurlyeq \operatorname{diam}(\omega_z)\big[\|\nabla\varphi^-\|_{\omega_z} + \|\nabla u_k\|_{\omega_z} + \mathcal{L}^d(\omega_z)^{1/2}M\big]$$

by using Lemma 3.2 for the first term. Combining this with (3.26) for $\zeta = \varphi^-$, we obtain

$$\rho_z(\varphi^-) \preccurlyeq \sum_{S\in\mathcal{S}_z\cap\mathcal{S}_k^1} d_S|\rho_S| + \left(\sum_{S\in\mathcal{S}_z\cap\mathcal{S}_k^2} |\rho_S|^2 + \|h_k f\|_{\omega_z}^2\right)^{1/2} \tag{3.32}$$
$$\times \big[\|\nabla\varphi^-\|_{\omega_z} + \|\nabla u_k\|_{\omega_z} + \mathcal{L}^d(\omega_z)^{1/2}M\big].$$

7. Using (3.23) and (3.27), we insert (3.25), (3.29), (3.31), and (3.32) into (3.22). Moreover, we note that each element (side) is contained at most in $d+1$ (d) stars and that $\|\nabla\varphi^+\|^2 + \|\nabla\varphi^-\|^2 = \|\nabla\varphi\|^2 \le 1$. Thus, we finally arrive at

$$\langle -\mathcal{D}_k, \varphi\rangle \preccurlyeq \left[\sum_{S\in\mathcal{S}_k^1} d_S|\rho_S| + \left(\sum_{S\in\mathcal{S}_k^2} |\rho_S|^2\right)^{1/2} + \left(\sum_{T\in\mathcal{T}_k^+}\int_T h_k^2|f|^2\right)^{1/2}\right] \times \left[1 + \|\nabla u_k\|_{\Omega_k^\perp} + \mathcal{L}^d(\Omega_k^\perp)^{1/2}M\right],$$

where $\Omega_k^\perp := \bigcup_{z\in\mathcal{N}_k^\perp}\omega_z$ denotes the union of all stars associated with isolated proper contact nodes. The stability of $\{u_k\}_k$, which is proved in the following lemma, then finishes the proof. ∎

LEMMA 3.3 (stability). *The approximate minimizers $\{u_k\}_k$ are stable: there is a constant C depending on f, $\{\psi_i\}_{i=1}^n$, and $\mathcal{T}_0$ such that, for any admissible k,*

$$\|\nabla u_k\| \le C. \tag{3.33}$$

Proof. Let ψ denote the Lagrange interpolation of $\max\{0,\psi^*\}$ onto the initial finite element space V_0. Then $\psi\in\mathcal{F}_k$ for any k, and so we can choose $v=\psi$ in (2.5) and obtain (3.33) by standard manipulations. ∎

The combination of Propositions 3.1 and 3.2 yields the main result of this section.

THEOREM 3.1 (upper bound). *The error estimator $\mathcal{E}_k$ in* (2.14) *bounds the error $I[u_k]-I[u]$ in the energy minimum from above. More precisely, there holds*

$$I[u_k] - I[u] \preccurlyeq \max\{\tfrac12\mathcal{E}_k^2, \mathcal{E}_k\}.$$

The hidden constant depends on d and the shape regularity σ_0 in (2.2) *and, only if the set of isolated proper contact nodes $\mathcal{N}_k^\perp$ defined in* (3.21) *is not empty, in addition on the initial mesh $\mathcal{T}_0$, the load term f, and the obstacle functions $\{\psi_i\}_{i=1}^n$.*

Remark 3.2 (isolated proper contact nodes). The existence of isolated, proper contact nodes can be easily verified a posteriori. They appear on relatively coarse meshes, e.g., if the obstacle is a pyramid. On finer meshes, they do not persist and get proper contact nodes that are not isolated. Thus, at least in this generic situation, the hidden constant in Theorem 3.1 asymptotically depends only on d and σ_0.

Remark 3.3 (localization and stability). In view of the definition (2.13) of $\mathcal{T}_k^+$, the evaluation of the nonhierarchical part

$$\left(\sum_{T\in\mathcal{T}_k^+}\int_T h_k^2|f|^2\right)^{1/2}$$

might appear unstable because small changes in u_k may produce big changes in $\mathcal{T}_k^+$. However, in light of the results of Brezzi and Caffarelli [6], this occurs only when $|f|$ is small in the corresponding regions.

3.3. On efficiency. In this section we theoretically investigate the sharpness of the upper bound in Theorem 3.1, which is related to the efficiency of the stopping test (2.15).

We first consider the hierarchical part (2.16) of the estimator $\mathcal{E}_k$. By construction, its indicators satisfy the following local lower bounds with respect to an error in a "local" energy minimum ξ_S given by (2.8); see (2.10b):

$$\xi_S = \tfrac12\rho_S^2, \quad S\in\mathcal{S}_k^2, \qquad \text{and} \qquad \xi_S > \tfrac12 d_S|\rho_S|, \quad S\in\mathcal{S}_k^1. \tag{3.34}$$

Their global counterparts follow with the help of the convexity of the Lagrangian associated with I.

THEOREM 3.2 (lower bounds for hierarchical part of estimator). *The two sums of the hierarchical estimator part η_k from* (2.16) *bound the error $I[u_k] - I[u]$ in the energy minimum from below:*

$$\max\left\{\sum_{S\in\mathcal{S}_k^2}\rho_S^2,\ \sum_{S\in\mathcal{S}_k^1} d_S\,|\rho_S|\right\} \le 2(d+1)\big(I[u_k]-I[u]\big).$$

Proof. 1. For notational convenience, denote by $\bar{\mathcal{S}}_k$ all sides of $\mathcal{T}_k$, including the boundary ones, and define φ_S, ω_S also for $S\in\bar{\mathcal{S}}_k\setminus\mathcal{S}_k$. Let

$$\varphi = \frac{1}{d+1}\sum_{S\in\bar{\mathcal{S}}_k}\beta_S\varphi_S \tag{3.35}$$

be a linear combination of the test functions involved in η_k and observe that, on each element $T\in\mathcal{T}_k$, $u_k+\varphi$ is a convex combination of certain $v_S := u_k+\beta_S\varphi_S$, $S\in\bar{\mathcal{S}}_k$:

$$(u_k+\varphi)_{|T} = \frac{1}{d+1}\sum_{S\in\bar{\mathcal{S}}_k,\,S\subset T} v_{S|T}. \tag{3.36}$$

Thus, the convexity of $\mathbb{R}^d\ni p\mapsto\frac12|p|^2$ implies

$$\begin{aligned} I[u_k+\varphi] &= \sum_{T\in\mathcal{T}_k}\int_T \tfrac12|\nabla(u_k+\varphi)|^2 - f(u_k+\varphi)\\ &\le \frac{1}{d+1}\sum_{T\in\mathcal{T}_k}\sum_{S\in\bar{\mathcal{S}}_k,\,S\subset T}\int_T \tfrac12|\nabla v_S|^2 - fv_S. \end{aligned}$$

Subtracting this from $I[u_k] = \frac{1}{d+1}\sum_{T\in\mathcal{T}_k}\sum_{S\in\bar{\mathcal{S}}_k,\,S\subset T}\int_T\big(\frac12|\nabla u_k|^2-fu_k\big)$ and reorganizing the sum leads to

$$\begin{aligned} &I[u_k]-I[u_k+\varphi]\\ &\quad\ge \frac{1}{d+1}\sum_{S\in\bar{\mathcal{S}}_k}\left[\int_{\omega_S}\big(\tfrac12|\nabla u_k|^2-fu_k\big)-\int_{\omega_S}\big(\tfrac12|\nabla v_S|^2-fv_S\big)\right]\\ &\quad\ge \frac{1}{d+1}\sum_{S\in\bar{\mathcal{S}}_k}\big(I[u_k]-I[u_k+\beta_S\varphi_S]\big). \end{aligned} \tag{3.37}$$

2. We first bound the sum associated with $\mathcal{S}_k^2$, and therefore we choose

$$\beta_S = \begin{cases}\alpha_S & \text{if } S\in\mathcal{S}_k^2,\\ 0 & \text{otherwise}\end{cases}\qquad \text{with } \alpha_S \text{ from (2.9)}$$

in (3.35). Since $\alpha_S\ge -d_S$, (2.7) implies $u_k+\alpha_S\varphi_S\in\mathcal{F}$ for all $S\in\mathcal{S}_k$. The convexity of $\mathcal{F}$ in combination with (3.36) then yields $u_k+\varphi\in\mathcal{F}$. For all interior sides $S\in\mathcal{S}_k$ the identity $I[u_k]-I[u_k+\alpha_S\varphi_S]=\xi_S=\frac12\rho_S^2$ holds by (2.8) and (3.34). With the help of (3.37), we thus obtain

$$I[u_k]-I[u]\ge I[u_k]-I[u_k+\varphi] = \frac{1}{d+1}\sum_{S\in\mathcal{S}_k^2}\xi_S = \frac{1}{2(d+1)}\sum_{S\in\mathcal{S}_k^2}\rho_S^2.$$

The sum over $\mathcal{S}_k^1$ is bounded analogously. ∎

We next discuss the efficiency of the estimator part related to the load f and then finish the section by confronting Theorems 3.1 and 3.2.

Remark 3.4 (lower bounds for nonhierarchical estimator part). Introducing φ_T, x_T, ρ_T, d_T, and ξ_T for elements $T \in \mathcal{T}_k^+$ in an analog manner as for sides, one can prove that, on unconstrained elements $T \in \mathcal{T}_k^2 := \{T \in \mathcal{T}_k^+ \mid \rho_T \geq -d_T\}$, there hold the local lower bounds

$$\int_T h_k^2 |f|^2 \preccurlyeq \xi_T + \int_T h_k^2 |f - \bar{f}_T|^2$$

with $\bar{f}_T := \mathcal{L}^d(T)^{-1} \int_T f$ in the oscillation term. This implies the global lower bound

$$\sum_{T \in \mathcal{T}_k^2} \int_T h_k^2 |f|^2 \preccurlyeq I[u_k] - I[u] + \sum_{T \in \mathcal{T}_k^2} \int_T h_k^2 |f - \bar{f}_T|^2;$$

that is, the nonhierarchical part of the estimator is sharp apart from a higher order term and the contribution associated with $\mathcal{T}_k^1 := \mathcal{T}_k^+ \setminus \mathcal{T}_k^2$. The latter is typically also of higher order because it is localized to a narrowing stripe around the exact free boundary. We omit the proof since this property is not used below.

As in the unconstrained problem, it is also possible to derive an upper bound that deals also with the load term in a hierarchical way apart from a higher oscillation part. Since the cost for computing such an estimator is higher (especially for $d = 3$), we used the estimator given in (2.14).

Remark 3.5 (gap between upper and lower bound). The powers appearing in Theorems 3.1 and 3.2 do not match, thus producing a gap between upper and lower bounds; e.g., the asymptotically relevant one corresponds to the second case in Proposition 3.1 and amounts to

$$I[u_k] - I[u] \preccurlyeq \mathcal{E}_k \preccurlyeq \big(I[u_k] - I[u]\big)^{1/2}$$

for $f = 0$ and $I[u_k] - I[u] \leq 1$. The reason for the gap is that, in Proposition 3.1, we distinguish an "unconstrained" and a "constrained" case in a global manner, while in Theorem 3.2, the cases are distinguished in a local manner, which may mismatch with the global choice. Notice, however, that our numerical experiments indicate a linear estimator-error relationship $\mathcal{E}_k \approx \big(I[u_k] - I[u]\big)^{1/2}$ as in the unconstrained problem; see, e.g., section 5.2.

4. Convergence of approximate minima. The goal of this section is to prove that the algorithm described in section 2 converges for $\mathsf{tol} = 0$. Since the algorithm depends on tol only through the stopping test (2.15), this implies in particular that, for any strictly positive tolerance $\mathsf{tol} > 0$, the algorithm finishes after a finite number of steps. Notice that the algorithm of section 2 does not ensure that the maximum meshsize $\max_\Omega h_k$ decreases to 0. In fact, this may not happen; see, e.g., section 5.1. Therefore, the convergence theory for nonadaptive finite elements does not apply.

THEOREM 4.1 (convergence). *Suppose that* $\mathsf{tol} = 0$ *and that the initial triangulation* $\mathcal{T}_0$ *is subordinated to the lower obstacle in the sense of* (2.1).

Then the algorithm of section 2 *converges in a finite number of steps or produces an infinite sequence of approximate minima* $\{u_k\}_{k \in \mathbb{N}}$ *such that*

$$I[u_k] \to I[u] \quad \textit{and} \quad u_k \to u \textit{ in } H^1(\Omega) \quad (k \to \infty).$$

Proof. 1. If the algorithm finishes after a finite number of steps, the upper bound in Theorem 3.1 guarantees that the last approximate minimizer coincides with the exact one.

Suppose that the algorithm never finishes. Then (2.22) ensures that

$$m := \lim_{k\to\infty} I[u_k]$$

exists, and we need to check that $m = I[u]$. In view of the characterization (1.3) of the exact minimizer and the property (3.6) of the seminorm ρ_k, we may do this by showing

$$\lim_{k\to\infty} \rho_k(-\mathcal{D}_k) = 0. \tag{4.1}$$

To this end, our starting point is the convergence of the energy reduction in one adaptive iteration,

$$\lim_{k\to\infty} \big(I[u_k] - I[u_{k+1}]\big) = 0, \tag{4.2}$$

which is an immediate consequence of the existence of $\lim_{k\to\infty} I[u_k]$.

2. As a first step towards (4.1), we show that the hierarchical part (2.16) of the estimator tends to 0, i.e., $\lim_{k\to\infty} \eta_k = 0$. The idea is, exploiting (2.20), to modify the proof of the lower bounds in Theorem 3.2. Let us choose

$$\beta_S = \begin{cases} \alpha_S & \text{if } S \in \hat{\mathcal{S}}_k \cap \mathcal{S}_k^2, \\ 0 & \text{otherwise} \end{cases} \qquad \text{with } \alpha_S \text{ from (2.9)}$$

in (3.35). Then (2.20) implies $\varphi_S \in V_{k+1}$ for all $S \in \hat{\mathcal{S}}_k$, and thus we conclude $u_k + \varphi \in \mathcal{F}_{k+1}$ with the same arguments used in the proof of Theorem 3.2. By (3.37) we then obtain

$$\sum_{S\in\hat{\mathcal{S}}_k\cap\mathcal{S}_k^2} \rho_S^2 \le 2(d+1)\big(I[u_k] - I[u_k+\varphi]\big) \le 2(d+1)\big(I[u_k] - I[u_{k+1}]\big).$$

Analogously we derive

$$\sum_{S\in\hat{\mathcal{S}}_k\cap\mathcal{S}_k^1} d_S|\rho_S| \le 2(d+1)\big(I[u_k] - I[u_{k+1}]\big).$$

Combining these two inequalities with the selection criterion (2.17a) for $\hat{\mathcal{S}}_k$, we obtain

$$\eta_k \le \frac{1}{\theta}\left[\left(\sum_{S\in\hat{\mathcal{S}}_k\cap\mathcal{S}_k^2} \rho_S^2\right)^{1/2} + \sum_{S\in\hat{\mathcal{S}}_k\cap\mathcal{S}_k^1} d_S|\rho_S|\right] \le g\big(I[u_k] - I[u_{k+1}]\big) \tag{4.3}$$

with $g(t) := 4\theta^{-1}(d+1)\max\{\sqrt{t}, t\}$ for all $t \ge 0$. Thanks to (4.2), we thus devise

$$\lim_{k\to\infty} \eta_k = 0. \tag{4.4}$$

3. Next, we verify the convergence of the nonhierarchical estimator part with the help of (2.17b) and (2.19b). We may write

$$\sum_{T\in\mathcal{T}_k^+} \int_T h_k^2|f|^2 = \int_\Omega h_k^2|f|^2\chi_k,$$

where $\chi_k(x) = 1$ if there is a $T \in \mathcal{T}_k^+$ with $T \ni x$ and $\chi_k(x) = 0$ otherwise. In view of Lebesgue's theorem about dominated convergence, it is sufficient to show that the integrand $h_k^2|f|^2\chi_k$ tends to 0 a.e. in Ω. Note that the set $\Gamma_\infty = \bigcup\{\partial T : T \in \mathcal{T}_k,\ k \in \mathbb{N}\}$ has Lebesgue measure 0 and that, for each $x \in \Omega \setminus \Gamma_\infty$, the sequence $\{h_k(x)\}_{k\in\mathbb{N}}$ is nonincreasing and nonnegative. Consequently, $h_\infty(x) := \lim_{k\to\infty} h_k(x)$ exists and is nonnegative for almost all $x \in \Omega$. Obviously, there holds

$$\lim_{k\to\infty} h_k^2|f|^2\chi_k = 0 \text{ a.e. in } \{h_\infty = 0\} \tag{4.5}$$

with $\{h_\infty = 0\} := \{x \in \Omega \mid h_\infty(x) = 0\}$. In view of (2.19b), the remaining set $\{h_\infty > 0\}$ is covered by the elements of $\mathcal{T}_* := \bigcup_{k\in\mathbb{N}} \bigcap_{\ell\geq k} \mathcal{T}_\ell$ that, after a certain step, will never be refined. Let $T \in \mathcal{T}_*$ be arbitrary and consider two cases.

Case 1: There exists $k \in \mathbb{N}$ such that $T \notin \mathcal{T}_\ell^+$ for all $\ell > k$. Then we readily obtain $\lim_{\ell\to\infty} h_\ell^2|f|^2\chi_\ell = 0$ a.e. in T from $\chi_\ell = 0$ on T for all $\ell > k$.

Case 2: For all $k \in \mathbb{N}$ exists $\ell > k$ with $T \in \mathcal{T}_\ell^+$. The fact that $T \in \mathcal{T}_*$ is not refined after a certain step implies the existence of a subsequence $\{\mathcal{T}_{k_\ell}\}_{\ell\in\mathbb{N}}$ of $\{\mathcal{T}_k\}_{k\in\mathbb{N}}$ such that $T \in \mathcal{T}_{k_\ell}^+ \setminus \hat{\mathcal{T}}_{k_\ell}$ for all $\ell \in \mathbb{N}$. The selection criterion (2.17b) then yields

$$\int_T h_\infty^2|f|^2 = \int_T h_{k_\ell}^2|f|^2 \leq \frac{\eta_{k_\ell}^2}{N_{k_\ell}^+} \leq \eta_{k_\ell}^2,$$

which in turn entails $f = 0$ a.e. in T thanks to (4.4) and $h_{\infty|T} > 0$. Consequently, we obtain $\lim_{k\to\infty} h_k^2|f|^2\chi_k = 0$ a.e. in T also in this case.

Combining the two cases, taking into account that $\mathcal{T}_*$ is separable, and invoking (4.5) yields

$$\lim_{k\to\infty} h_k^2|f|^2\chi_k = 0 \text{ a.e. in } \Omega.$$

Hence, the dominated convergence theorem implies

$$\lim_{k\to\infty} \left(\sum_{T\in\mathcal{T}_k^+} \int_T h_k^2|f|^2 \right)^{1/2} = 0. \tag{4.6}$$

4. Inserting (4.4) and (4.6) into (2.14) yields $\lim_{k\to\infty} \mathcal{E}_k = 0$. Thanks to Proposition 3.2 and Theorem 3.1, we thus obtain $\lim_{k\to\infty} \rho_k(-\mathcal{D}_k) = 0$ as well as

$$m = \lim_{k\to\infty} I[u_k] = I[u].$$

The convergence $u_k \to u$ in $H^1(\Omega)$ then follows from (3.4) and the well-known Poincaré inequality $\|u\| \leq C_\Omega \|\nabla u\|$. □

Remark 4.1 (strict decrease). The estimate (4.3) in step 2 shows that there holds $I[u_{k+1}] < I[u_k]$ whenever the hierarchical part η_k of the estimator is nonzero.

5. Numerical results. In this section we present a couple of examples implemented within the finite element toolbox ALBERTA of Schmidt and Siebert [20, 21]. The computation of the hierarchical estimator involves a global refinement of $\mathcal{T}_k$ that contains for all sides $S \in \mathcal{S}_k$ the corresponding nodes x_S. This globally refined mesh contains the mesh $\mathcal{T}_{k+1}$ of the next step but is in general larger, because not all elements in $\mathcal{T}_k$ are refined when creating $\mathcal{T}_{k+1}$. In order to obtain a cheap realization

of the global refinement of $\mathcal{T}_k$, we correspondingly refine all elements of $\mathcal{T}_k$ in a separate and virtual manner. More precisely, with the help of the rules for refinement by bisection, each element is virtually refined without using information on neighbors and without changing the data structures of the current mesh $\mathcal{T}_k$. The contributions of the hierarchical estimator are then computed on the virtual subelements. If an element is actually refined later on (e.g., when producing $\mathcal{T}_{k+1}$), it is refined in the same way it was virtually refined.

For the virtual refinement of a single element, we use refinement patterns for reference situations that ensure the additional nodes for all sides. In two dimensions, the situation is rather simple, and we need only the one refinement pattern depicted in Figure 3, which exactly corresponds to all children of the second generation. In three dimensions, the situation is more involved: There are elements of three different types where the element type is uniquely assigned by the refinement procedure; see [21, section 1.1.1]. The creation of all children of the fifth generation is sufficient to ensure interior nodes for all sides, but these nodes are already created with fewer bisections producing a certain mixture of children of the fourth and fifth generations. In order to minimize work in the computation of the estimator, we have stored for each element type a refinement pattern with the fewest number of children that guarantees the additional nodes for all faces. However, for the sake of simplicity of the refinement of $\mathcal{T}_k$ into $\mathcal{T}_{k+1}$, we create in (2.18) all children of fifth generation. Note that the test functions φ_S created by the virtual refinement can be written as linear combinations of few hat functions in V_{k+1}.

The estimator is implemented in the form $\mathcal{E}_k := \mathcal{E}_{k,1} + \mathcal{E}_{k,2} + \mathcal{E}_{k,3}$, where

$$\mathcal{E}_{k,1} := c_1 \sum_{S\in\mathcal{S}_k^1} d_S|\rho_S|, \quad \mathcal{E}_{k,2} := c_2 \left(\sum_{S\in\mathcal{S}_k^2} \rho_S^2 \right)^{1/2}, \quad \mathcal{E}_{k,3} := c_3 \left(\sum_{T\in\mathcal{T}_k^+} \int_T h_k^2 |f|^2 \right)^{1/2}$$

and the constants are chosen in all experiments as $c_1 = c_2 = 0.25$ and $c_3 = 0.5$.

5.1. Effect of localization for a Lipschitz obstacle. In order to illustrate the effect of the localization of the estimator $\mathcal{E}_k$ to nonfull-contact regions, we consider an example with Lipschitz obstacle as in [18, section 7.4] and [19, section 3.2]. On the diamond domain $\Omega := \{x = (x_1, x_2) \mid |x_1| + |x_2| < 1\}$, let

$$\psi(x) := \operatorname{dist}(x, \partial\Omega) - \tfrac{1}{2},$$

and $f(x) := -15$. The graph of ψ is a pyramid with a square base centered at the origin; see Figure 4 (left), where also the initial mesh is depicted. Hence, due to the relatively heavy load, the exact solution is in contact in the middle of the domain, then detaches, and finally increases towards the boundary; see Figure 4 (right) for the finite element minimizer u_5 of iteration $k = 5$. Figure 5 displays the meshes for iterations $k \in \{4, 5, 6\}$. Notice that they remain coarse in the contact region, even where the gradient of the finite element minimizers jumps. This saving of degrees of freedoms (DOFs), which is also helpful for solving (2.5) in less stable situations, is thanks to the localization of $\mathcal{E}_k$ to the nonfull-contact region.

5.2. Estimator effectivity for stable free boundaries in two and three dimensions. Next, in the situation of stable free boundaries, we examine the error-estimator relationship and the importance of $\mathcal{E}_{k,1}$—the estimator part that is built up with the "constrained" sides $\mathcal{S}_k^1$. On the domains $\Omega = (-1, 1)^d$ with $d \in \{2, 3\}$,

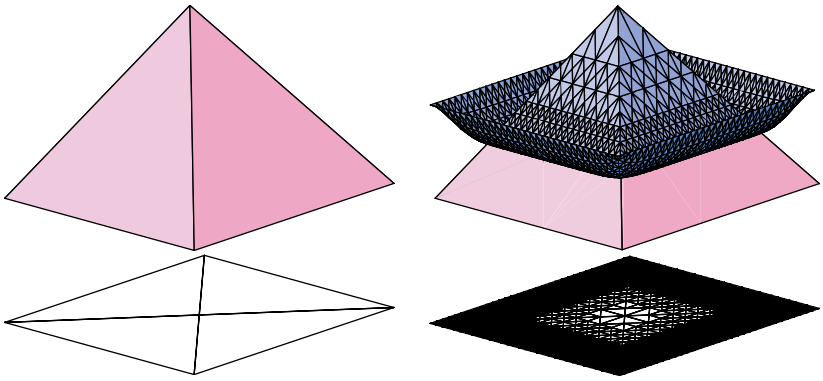


FIG. 4. *Lipschitz obstacle: graphs of the obstacle with initial mesh (left) and finite element minimizer u_5 with corresponding mesh $\mathcal{T}_5$ (right).*

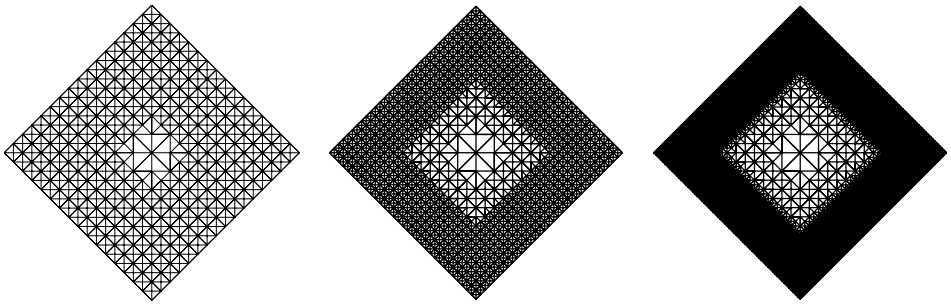


FIG. 5. *Lipschitz obstacle: meshes $\mathcal{T}_k$ for $k \in \{4,5,6\}$. The meshes remain coarse in the contact region thanks to the localization of $\mathcal{E}_k$ to the nonfull-contact region.*

consider a constant obstacle $\psi(x) := 0$, the continuous load

$$f(x) := \begin{cases} -4\left(2\,|x|^2 + d\,(|x|^2 - r^2)\right), & |x| > r, \\ -8\,r^2\left(1 - (|x|^2 - r^2)\right), & |x| \le r, \end{cases}$$

and the Dirichlet boundary values $g(x) := (|x|^2 - r^2)^2$ with $r = 0.7$. (We neglect the approximation of the boundary data that is not piecewise affine.) The exact minimizer to this problem is

$$u(x) = \left(\max\{|x|^2 - r^2, 0\}\right)^2;$$

see Figure 6 for a finite element minimizer in the case $d = 2$. Figure 7 depicts the estimator $\mathcal{E}_k$ from (2.14), the square root of the error $\left(I[u_k] - I[u]\right)^{1/2}$ in the energy minimum, and the energy norm error $\|\nabla(u_k - u)\|$ versus the number of DOFs in a log-log scale. We can see that, at least in these cases, the estimator $\mathcal{E}_k$ is equivalent with $\left(I[u_k] - I[u]\right)^{1/2}$ and $\|\nabla(u_k - u)\|$. Table 5.1 reveals that the constrained part $\mathcal{E}_{k,1}$ is clearly dominated by the unconstrained part $\mathcal{E}_{k,2}+\mathcal{E}_{k,3}$, thus providing a partial explanation for the aforementioned equivalence. These two observations are valid also for the experiments of the following sections, which cover unstable free boundaries. This indicates that, at least typically, the stopping test (2.15) is reliable and efficient

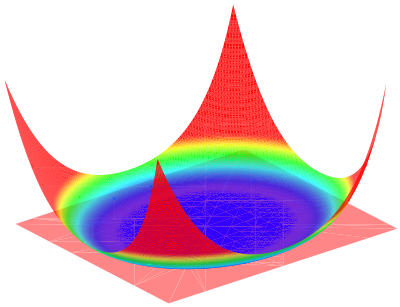


FIG. 6. *Stable free boundary: a finite element minimizer with obstacle for* $d = 2$.

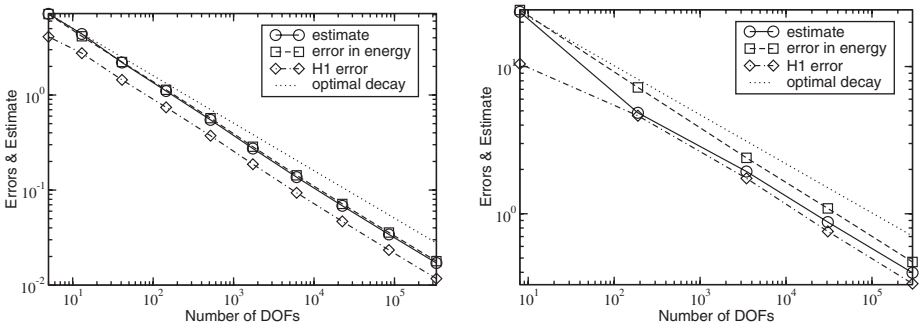


FIG. 7. *Stable free boundary: estimator, error in energy minimum, and energy norm error versus DOFs in log-log scale for* $d = 2$ *(left) and* $d = 3$ *(right). Estimator and errors are equivalent, and their asymptotic decays coincide with the best possible decay rate* $-1/2$ *and* $-1/3$ *for linear finite elements in two and three dimensions, respectively.*

TABLE 5.1

Stable free boundary: comparison of the estimator contributions for $d = 2$ *(left) and* $d = 3$ *(right). The constrained part* $\mathcal{E}_{k,1}$ *is clearly dominated by the unconstrained part* $\mathcal{E}_{k,2} + \mathcal{E}_{k,3}$.

DOFs	$\mathcal{E}_{k,1}$	$\mathcal{E}_{k,2} + \mathcal{E}_{k,3}$
5	0.000e+00	7.152e+00
13	2.579e−01	4.386e+00
41	0.000e+00	2.197e+00
145	1.432e−03	1.101e+00
517	2.232e−04	5.478e−01
1737	1.877e−05	2.734e−01
6077	3.722e−06	1.366e−01
22461	3.353e−07	6.828e−02
85365	3.447e−08	3.413e−02
332245	2.665e−09	1.706e−02

DOFs	$\mathcal{E}_{k,1}$	$\mathcal{E}_{k,2} + \mathcal{E}_{k,3}$
8	0.000e+00	2.352+01
189	0.000e+00	4.850+00
3465	2.150e−04	1.936+00
30667	6.043e−06	8.777−01
296495	1.061e−06	3.970−01

in the usual linear sense; i.e., there holds $\mathcal{E}_k \approx \big(I[u_k] - I[u]\big)^{1/2} \approx \|\nabla(u_k - u)\|$; see also Remarks 3.1 and 3.5.

5.3. Insensitivity to ineffective data changes. The following examples serve to investigate the behavior of the adaptive algorithm with respect to changes in the data that are not reflected in the exact solution. Recall that the adaptive algorithm aims only at the approximation of the exact minimizer and therefore should be insensitive to such changes.

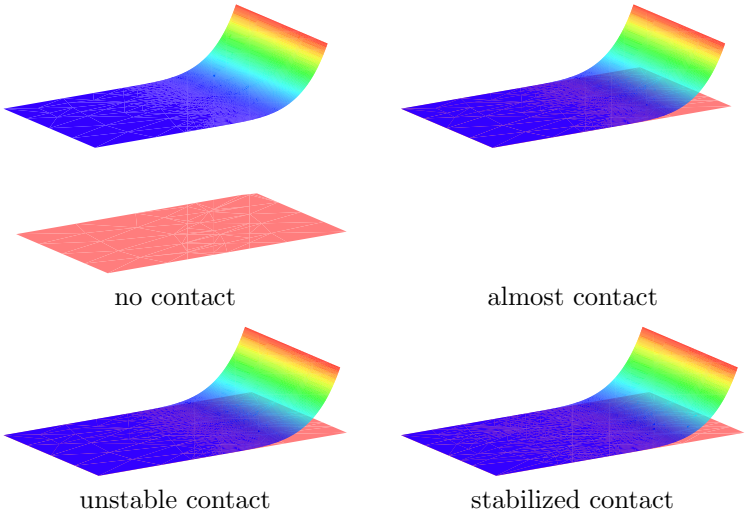


FIG. 8. *Ineffective data changes: comparison of finite element minimizers that approximate the same solution for different data ranging from "no contact" to "stabilized contact." There is no visual difference between the discrete minimizers.*

We consider four examples that all have the same exact solution,

$$u(x_1,x_2) := \begin{cases} \frac{1}{2}x_1^4 & \text{if } x_1 \geq 0, \\ 0 & \text{else} \end{cases} \qquad \text{for } (x_1,x_2) \in \Omega := (-1,1)^2,$$

but different obstacles and loads so that this solution is

- unconstrained,
- "almost in contact" with the obstacle,
- just and thus in a "unstable" manner in contact, and
- in "stabilized contact."

More precisely, let

$$\psi_1 \equiv -1, \quad \psi_2 \equiv -0.001, \quad \psi_3 \equiv \psi_4 \equiv 0$$

and

$$f_i(x_1,x_2) := \begin{cases} -6x_1^2 & \text{if } x_1 \geq 0, \\ 0 & \text{if } x_1 \leq 0 \text{ and } i \in \{1,2,3\}, \\ x_1 & \text{if } x_1 \leq 0 \text{ and } i = 4. \end{cases}$$

The initial mesh is the same for all four examples and not aligned with the one-dimensional structure of the exact solution. Figure 8 depicts for all four examples the finite element minimizers u_4 and Figure 9 the corresponding meshes $\mathcal{T}_4$, while Figure 10 shows the decays of estimator and errors versus number of DOFs in a log-log scale.

We notice that the finite element minimizers and their meshes do not depend, or only little, on the four different data regimes. The same holds true for the decays and, moreover, the single estimator contributions; see Table 5.2. This indicates that the estimator and adaptive algorithm depend essentially only on the approximated exact solution. In particular, it appears insensitive to contact set changes that are not reflected in the solution, even if they are discontinuous; see also Remark 3.3.

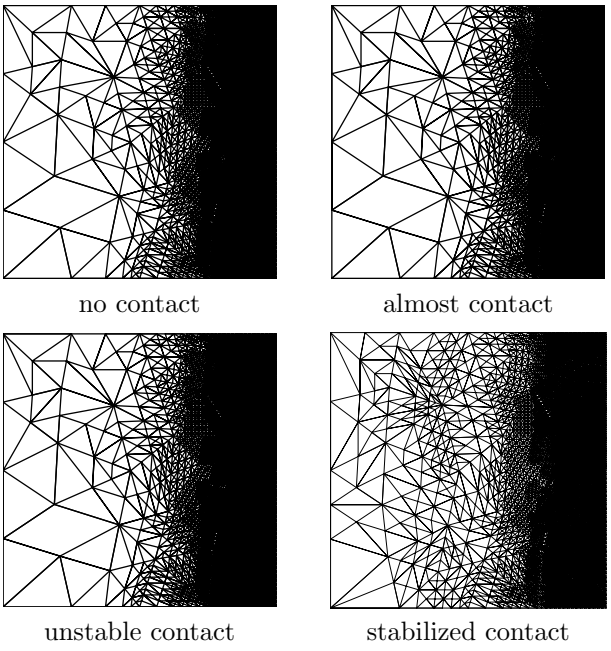


FIG. 9. *Ineffective data changes: comparison of the meshes $\mathcal{T}_4$. Only in the last case of stabilized contact is there a slight difference around the free boundary $x_1 = 0$.*

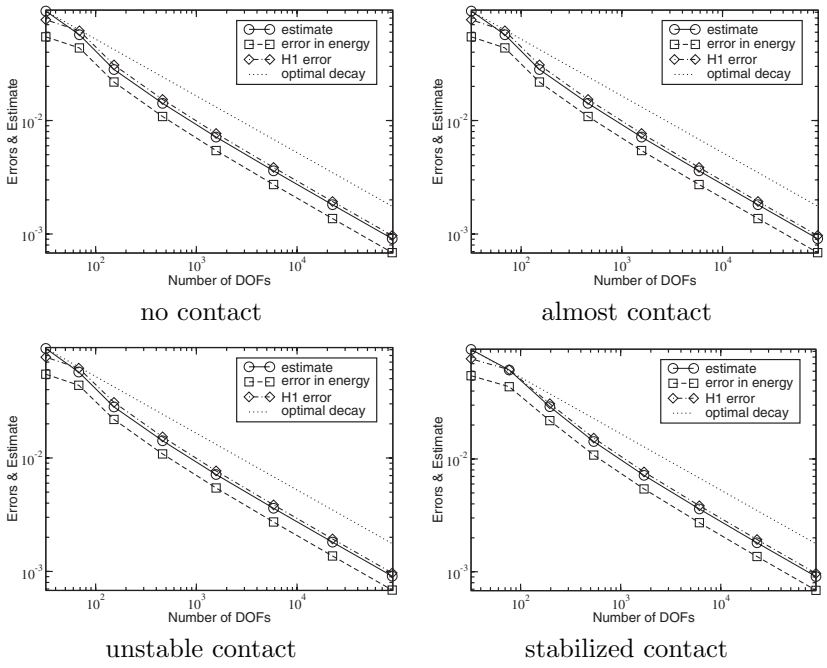


FIG. 10. *Ineffective data changes: comparison of convergence histories of estimator, error in energy minimum, and energy norm error versus number of DOFs in log-log scale. The histories are essentially independent of the varying data, and their asymptotic decays coincide with the best possible decay rate $-1/2$ for linear finite elements.*

TABLE 5.2

Ineffective data changes: comparison of estimator contributions from "no contact" (top) to "stabilized contact" (bottom). Single contributions are essentially independent of varying data.

DOFs	$\mathcal{E}_{k,1}$	$\mathcal{E}_{k,2}$	$\mathcal{E}_{k,3}$
32	0.000e+00	1.232e−01	4.298e−01
69	0.000e+00	1.269e−01	2.151e−01
152	0.000e+00	6.020e−02	1.088e−01
462	0.000e+00	2.935e−02	5.567e−02
1571	0.000e+00	1.469e−02	2.823e−02
5841	0.000e+00	7.381e−03	1.424e−02
22487	0.000e+00	3.704e−03	7.162e−03
88086	0.000e+00	1.857e−03	3.595e−03

DOFs	$\mathcal{E}_{k,1}$	$\mathcal{E}_{k,2}$	$\mathcal{E}_{k,3}$
32	7.802e−08	1.240e−01	4.298e−01
69	1.380e−06	1.276e−01	2.151e−01
152	0.000e+00	6.020e−02	1.088e−01
462	0.000e+00	2.935e−02	5.567e−02
1571	0.000e+00	1.469e−02	2.823e−02
5841	0.000e+00	7.381e−03	1.424e−02
22487	0.000e+00	3.704e−03	7.162e−03
88086	0.000e+00	1.857e−03	3.595e−03

DOFs	$\mathcal{E}_{k,1}$	$\mathcal{E}_{k,2}$	$\mathcal{E}_{k,3}$
32	0.000e+00	1.242e−01	4.298e−01
68	5.559e−07	1.276e−01	2.152e−01
152	1.620e−08	6.020e−02	1.088e−01
462	0.000e+00	2.935e−02	5.567e−02
1571	0.000e+00	1.469e−02	2.823e−02
5841	0.000e+00	7.381e−03	1.424e−02
22487	0.000e+00	3.704e−03	7.162e−03
88086	0.000e+00	1.857e−03	3.595e−03

DOFs	$\mathcal{E}_{k,1}$	$\mathcal{E}_{k,2}$	$\mathcal{E}_{k,3}$
32	0.000e+00	1.242e−01	4.300e−01
68	1.149e−06	1.276e−01	2.174e−01
162	8.153e−07	6.022e−02	1.106e−01
479	4.393e−07	2.943e−02	5.662e−02
1635	8.578e−08	1.472e−02	2.840e−02
5953	9.644e−10	7.384e−03	1.427e−02
22661	7.813e−12	3.708e−03	7.166e−03
88268	6.298e−12	1.857e−03	3.596e−03

5.4. Optimal convergence speed for a discontinuous and thin obstacle. The main features of our last example are the singularities of the exact solution that are induced by the discontinuous and thin nature of the obstacle. Figure 11 (left) depicts a finite element minimizer in $\Omega = (-1,1) \times (-\frac{1}{2},\frac{1}{2})$ and reveals also the obstacle that is made of two parts: a "triangle-shaped" one with constant level, and a "segment-shaped" one with increasing level; to be more precise,

$$K_1 := \text{conv hull}\{(0,0),\, (\tfrac{1}{2},0),\, (\tfrac{1}{2},\tfrac{1}{4})\}, \qquad \psi_1(x_1,x_2) = 1,$$
$$K_2 := \text{conv hull}\{(-1,\tfrac{1}{2}),\, (-\tfrac{1}{2},\tfrac{1}{4})\}, \qquad \psi_2(x_1,x_2) = 2 + 2x_1.$$

The load term f is set to 0. Obviously, the induced singularities are related to the ones of solutions to Poisson's equation on domains with reentrant corners or on the crack domain. The solution u is at most in $H^{3/2}(\Omega)$, and thus, for nonadaptive

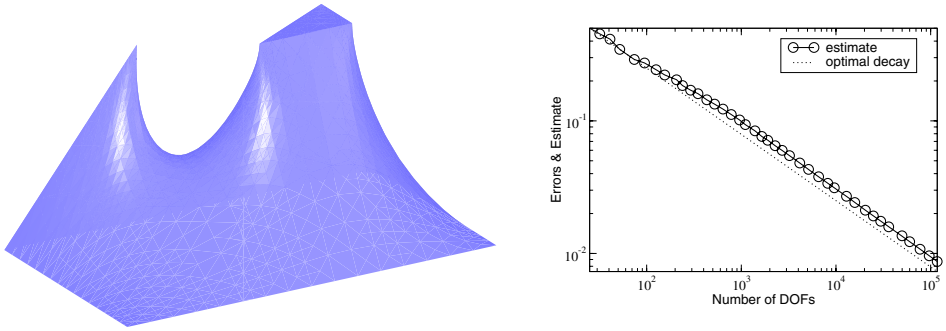


FIG. 11. *Discontinuous and thin obstacle: a finite element minimizer (left) and decay of estimator versus number of DOFs, together with maximum decay rate for nonlinear approximation with linear finite elements, in log-log scale (right).*

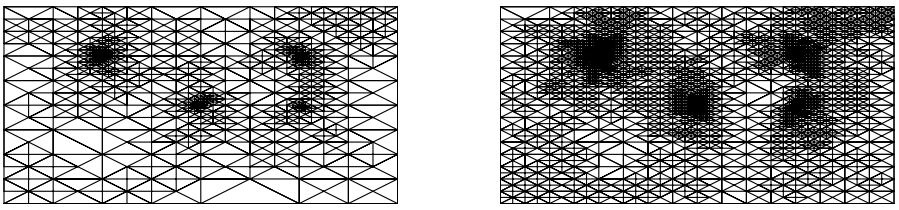


FIG. 12. *Discontinuous and thin obstacle: adaptive meshes $\mathcal{T}_k$ for iteration $k = 15$ (left) and $k = 22$ (right).*

uniform refinement, the decay rate of the energy norm error with respect to the number of DOFs is not better than $-1/4$, see, e.g., [7, Ch. 23]. However, the second derivatives of u are better than $L_1(\Omega)$, entailing that u in the energy norm error can be approximated on adaptive meshes with rate $-1/2$, which is the best decay rate that can be attained with linear finite elements for $d = 2$; see [5, Theorem 9.1 and Remark 9.2].

Figure 12 depicts two correspondingly adapted meshes, and Figure 11 (right) indicates the decay of the estimator $\mathcal{E}_k$ versus the number of DOFs in a log-log scale. The numerically observed decay rate of the estimator is $-1/2$. Notice that in the present case our theory ensures

$$\tfrac{1}{2}\|\nabla(u_k - u)\|^2 \leq I[u_k] - I[u] \approx \tfrac{1}{2}\mathcal{E}_k^2. \tag{5.1}$$

Indeed, the first inequality is just (3.4) and "$\approx$" holds because, on one hand, (3.13) applies, see Remark 3.1, while on the other hand, we need only estimate the term with the unconstrained sides $\mathcal{S}_k^2$ in Theorem 3.2 since the constrained sides $\mathcal{S}_k^1$ in Proposition 3.2 do not contribute. Consequently, (5.1) and the observation that the estimator decays with the best possible rate for the energy norm error mean that the estimator, the error in the energy minimum, and the energy norm error decay with the optimal rate.

Acknowledgment. We thank an anonymous referee for various comments that improved the presentation.

REFERENCES

[1] M. Ainsworth and J. T. Oden, *A Posteriori Error Estimation in Finite Element Analysis*, Pure Appl. Math. (NY), Wiley-Interscience, John Wiley and Sons, New York, 2000.

[2] I. Babuška and T. Strouboulis, *The Finite Element Method and Its Reliability*, Numer. Math. Sci. Comput., The Clarendon Press, Oxford University Press, New York, 2001.

[3] E. Bänsch, P. Morin, and R. H. Nochetto, *An adaptive Uzawa FEM for the Stokes problem: Convergence without the inf-sup condition*, SIAM J. Numer. Anal., 40 (2002), pp. 1207–1229.

[4] S. Bartels and C. Carstensen, *Averaging techniques yield reliable a posteriori finite element error control for obstacle problems*, Numer. Math., 99 (2004), pp. 225–249.

[5] P. Binev, W. Dahmen, and R. DeVore, *Adaptive finite element methods with convergence rates*, Numer. Math., 97 (2004), pp. 219–268.

[6] F. Brezzi and L. A. Caffarelli, *Convergence of the discrete free boundaries for finite element approximations*, RAIRO Anal. Numér., 17 (1983), pp. 385–395.

[7] P. G. Ciarlet, *Basic error estimates for elliptic problems*, in Handbook of Numerical Analysis, Vol. II, P. G. Ciarlet and J.-L. Lions, eds., North–Holland, Amsterdam, 1991, pp. 17–352.

[8] S. Dahlke, R. Hochmuth, and K. Urban, *Adaptive wavelet methods for saddle point problems*, M2AN Math. Model. Numer. Anal., 34 (2000), pp. 1003–1022.

[9] W. Dörfler, *A convergent adaptive algorithm for Poisson's equation*, SIAM J. Numer. Anal., 33 (1996), pp. 1106–1124.

[10] C. M. Elliott, *On the finite element approximation of an elliptic variational inequality arising from an implicit time discretization of the Stefan problem*, IMA J. Numer. Anal., 1 (1981), pp. 115–125.

[11] F. Fierro and A. Veeser, *A posteriori error estimators for regularized total variation of characteristic functions*, SIAM J. Numer. Anal., 41 (2003), pp. 2032–2055.

[12] D. Kinderlehrer and G. Stampacchia, *An Introduction to Variational Inequalities and Their Applications*, Pure Appl. Math. 88, Academic Press, New York, 1980.

[13] R. Kornhuber, *A posteriori error estimates for elliptic variational inequalities*, Comput. Math. Appl., 31 (1996), pp. 49–60.

[14] R. Kornhuber, *Adaptive Monotone Multigrid Methods for Nonlinear Variational Problems*, Adv. Numer. Math., B. G. Teubner, Stuttgart, 1997.

[15] P. Morin, R. H. Nochetto, and K. G. Siebert, *Data oscillation and convergence of adaptive FEM*, SIAM J. Numer. Anal., 38 (2000), pp. 466–488.

[16] P. Morin, R. H. Nochetto, and K. G. Siebert, *Convergence of adaptive finite element methods*, SIAM Rev., 44 (2002), pp. 631–658.

[17] P. Morin, R. H. Nochetto, and K. G. Siebert, *Local problems on stars: A posteriori error estimators, convergence, and performance*, Math. Comp., 72 (2003), pp. 1067–1097.

[18] R. H. Nochetto, K. G. Siebert, and A. Veeser, *Pointwise a posteriori error control for elliptic obstacle problems*, Numer. Math., 95 (2003), pp. 163–195.

[19] R. H. Nochetto, K. G. Siebert, and A. Veeser, *Fully localized a posteriori error estimators and barrier sets for contact problems*, SIAM J. Numer. Anal., 42 (2005), pp. 2118–2135.

[20] A. Schmidt and K. G. Siebert, *ALBERT—Software for scientific computations and applications*, Acta Math. Univ. Comenian. (N.S.), 70 (2000), pp. 105–122.

[21] A. Schmidt and K. G. Siebert, *Design of Adaptive Finite Element Software: The Finite Element Toolbox ALBERTA*, Lect. Notes Comput. Sci. Engrg. 42, Springer, Berlin, 2005.

[22] L. R. Scott and S. Zhang, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.

[23] R. Stevenson, *Optimality of a standard adaptive finite element method*, Found. Comput. Math., published online, 2006, DOI 10.1007/s10208-005-0183-0.

[24] A. Veeser, *Efficient and reliable a posteriori error estimators for elliptic obstacle problems*, SIAM J. Numer. Anal., 39 (2001), pp. 146–167.

[25] A. Veeser, *Convergent adaptive finite elements for the nonlinear Laplacian*, Numer. Math., 92 (2002), pp. 743–770.

[26] R. Verfürth, *A Review of a Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Adv. Numer. Math., John Wiley, Chichester, UK, 1996.

SIAM J. OPTIM.
Vol. 18, No. 1, pp. 290–308

A VECTOR LABELING METHOD FOR SOLVING DISCRETE ZERO POINT AND COMPLEMENTARITY PROBLEMS*

GERARD VAN DER LAAN†, DOLF TALMAN‡, AND ZAIFU YANG§

Abstract. In this paper we establish the existence of a discrete zero point of a function from the n-dimensional integer lattice $\mathbb{Z}^n$ to the n-dimensional Euclidean space $\mathbb{R}^n$ under very general conditions with respect to the behavior of the function. The proof is constructive and uses a combinatorial argument based on a simplicial algorithm with vector labeling and lexicographic linear programming pivot steps. The algorithm provides an efficient method to find an exact solution. We also discuss how to adapt the algorithm for two related problems, namely, to find a discrete zero point of a function under a general antipodal condition and to find a solution to a discrete nonlinear complementarity problem. In both cases the modified algorithm provides a constructive existence proof, too. We further show that the algorithm for the discrete nonlinear complementarity problem generalizes the well-known Lemke's method to nonlinear environments. An economic application is also presented.

Key words. integer lattice, zero point, vector labeling rule, simplicial algorithm, discrete complementarity

AMS subject classifications. 47H10, 54H25, 55M20, 90C26, 90C33, 91B50

DOI. 10.1137/050646378

1. Introduction.

We consider the problem of finding a point $x^* \in \mathbb{Z}^n$ such that

$$f(x^*) = 0^n,$$

where 0^n is the n-vector of zeroes, $\mathbb{Z}^n$ is the integer lattice of the n-dimensional Euclidean space $\mathbb{R}^n$, and f is a function from $\mathbb{Z}^n$ to $\mathbb{R}^n$. Such an integral point x^* is called a *discrete zero point* of f. Recently, the existence problem of an integral solution has been investigated in several papers. These papers were all inspired by the discrete fixed point statement given in Iimura [11]. In Iimura, Murota, and Tamura [12] and Danilov and Koshevoy [4], the existence theorems concern functions that exhibit the so-called *direction-preserving property* proposed by Iimura [11], which can be seen as the counterpart of the continuity property for functions defined on the Euclidean space $\mathbb{R}^n$. The existence results in Yang [37] and [38] hold for the class of so-called *locally gross direction-preserving* mappings, which is substantially more general and richer than the class of Iimura's direction-preserving mappings and which contains the results in [4] and [12] as special cases. Besides establishing these more general existence results, Yang also initiated in [37] the study of discrete nonlinear complementarity problems and provided several general theorems for the existence of solutions for this class of problems. All this literature, however, is not concerned with the problem of finding an integral solution. In fact, all these existence proofs are nonconstructive.

*Received by the editors November 30, 2005; accepted for publication (in revised form) December 10, 2006; published electronically April 3, 2007.
http://www.siam.org/journals/siopt/18-1/64637.html
†Department of Econometrics and Tinbergen Institute, Vrije Universiteit, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands (glaan@feweb.vu.nl).
‡Department of Econometrics & Operations Research and CentER, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands (talman@uvt.nl).
§Faculty of Business Administration, Yokohama National University, Yokohama 240-8501, Japan (yang@ynu.ac.jp).

To provide constructive proofs based on a combinatorial argument, we apply the technique of the so-called simplicial algorithms originally designed to find approximate zero or fixed points of continuous functions or upper semicontinuous mappings. The first of such algorithms was developed by Scarf [28], and subsequent algorithms proposed by Eaves [5], Eaves and Saigal [6], Merrill [23], and van der Laan and Talman [17], among others, substantially improved Scarf's original algorithm in terms of efficiency and applicability. For comprehensive treatments on such algorithms we refer to Allgower and Georg [1], Todd [30], and Yang [36]. The $2n$-ray integer labeling algorithm in [18] and [26] has been modified by van der Laan, Talman, and Yang in [20] to find an integral zero point of a function satisfying the direction-preserving property and in [21] to find a solution of a discrete nonlinear complementarity problem.

The aim of this paper is to provide a combinatorial algorithm for finding an integral zero point of a function satisfying the more general simplicially local gross direction-preserving property. This algorithm is also a modification of the $2n$-ray simplicial algorithm introduced in [18] and [26]. However, in this case we cannot rely on integer labeling anymore; instead we have to apply the more subtle concept of vector labeling. The modified algorithm makes use of a triangulation of $\mathbb{R}^n$, being a family of integral simplices, constructed in such a way that the set of vertices of the simplices of the triangulation is equal to $\mathbb{Z}^n$ and the mesh size of each simplex in the triangulation is equal to 1 according to the maximum norm. Starting with some integral point in $\mathbb{Z}^n$, the algorithm leaves the starting point along one out of $2n$ directions and then generates a sequence of adjacent simplices of varying dimension by making lexicographic linear programming pivot steps in a system of linear equations. We show that under a mild convergence condition the algorithm ends in a finite number of steps with an exact integral zero point. It is worth mentioning that in the case of a continuous function on $\mathbb{R}^n$, algorithms for finding a zero (or fixed) point find only an approximate solution, whereas the current algorithm for the discrete case finds an exact solution.

We also discuss how to adapt the algorithm for two related problems, namely, to find a discrete zero point of a function under a general antipodal condition and to find a solution to a discrete nonlinear complementarity problem. In the first case the antipodal condition guarantees convergency; in the second case we also propose a convergence condition. We show that the modified algorithm for the discrete nonlinear complementarity problem generalizes the well-known Lemke's method. In particular, when the function $f(x)$ is affine, i.e., $f(x) = Mx + q$, where M is an $n \times n$ matrix and q is an integral n-vector, it is shown that the algorithm finds an integral solution provided that M is totally unimodular and copositive-plus, and the system of $Mx+q \geq 0^n$, $x \geq 0^n$, is feasible.

This paper is organized as follows. In section 2 we introduce the concepts of triangulation and simplicially local gross direction-preservingness and describe the algorithm. In section 3 we state a convergence condition guaranteeing the existence of an integral solution to the discrete zero point problem and provide a constructive proof. In section 4 we modify the algorithm for the case that the function satisfies a general antipodal condition. In section 5 we modify the algorithm for the discrete complementarity problem and show that this modified algorithm generalizes Lemke's method. An economic application is discussed in section 6.

2. A method for solving discrete nonlinear equations. For a given positive integer n, let N denote the set $\{1, 2, \ldots, n\}$. For $i \in N$, $e(i)$ denotes the ith unit vector of $\mathbb{R}^n$. Given a set $D \subset \mathbb{R}^n$, $\mathrm{Co}(D)$ and $\mathrm{Bd}(D)$ denote the convex hull of D and the relative boundary of D, respectively. For any x and y in $\mathbb{R}^n$, we say y is

lexicographically greater than x and denote it by $y \succeq x$, if the first nonzero component of $y - x$ is positive.

Two integral points x and y in $\mathbb{Z}^n$ are said to be *cell-connected* if $\max_{h \in N} |x_h - y_h| \leq 1$, i.e., their distance is less than or equal to 1 according to the maximum norm. In other words, two integral points x and y are cell-connected if and only if there exists $q \in \mathbb{Z}^n$ such that both x and y belong to the hypercube $[0,1]^n + \{q\}$.

For an integer t, $0 \leq t \leq n$, the t-dimensional convex hull of $t+1$ affinely independent points $x^1, \ldots, x^{t+1}$ in $\mathbb{R}^n$ is called a *t-simplex* or simply a *simplex* and will be denoted by $\langle x^1, \ldots, x^{t+1} \rangle$. The extreme points $x^1, \ldots, x^{t+1}$ of a t-simplex $\sigma = \langle x^1, \ldots, x^{t+1} \rangle$ are called the *vertices* of σ. The convex hull of any subset of $k+1$ vertices of a t-simplex σ, $0 \leq k \leq t$, is called a *face* or *k-face* of σ. A k-face of a t-simplex σ is called a *facet* of σ if $k = t-1$, i.e., if the number of vertices is just one less than the number of vertices of the simplex. A simplex is said to be *integral* if all of its vertices are integral vectors and are cell-connected. Any two vertices x and y of an integral simplex are said to be *simplicially connected*.

Given an m-dimensional convex set D, a collection $\mathcal{T}$ of m-dimensional simplices is a *triangulation* or *simplicial subdivision* of the set D if (i) D is the union of all simplices in $\mathcal{T}$, (ii) the intersection of any two simplices of $\mathcal{T}$ is either empty or a common face of both, and (iii) any neighborhood of any point in D meets only a finite number of simplices of $\mathcal{T}$. A facet of a simplex of $\mathcal{T}$ either lies on the boundary of D and is a facet of no other simplex of $\mathcal{T}$ or it is a facet of precisely one other simplex of $\mathcal{T}$. A triangulation is called *integral* if all its simplices are integral simplices. One of the most well-known integral triangulations of $\mathbb{R}^n$ is the K-triangulation owing to Freudenthal [8]. This triangulation is the collection of all integral simplices $\sigma(y, \pi)$ with vertices $y^1, \ldots, y^{n+1}$, where, for $y \in \mathbb{Z}^n$ and $\pi = (\pi(1), \ldots, \pi(n))$ a permutation of the elements $1, 2, \ldots, n$, the vertices are given by $y^1 = y$ and $y^{i+1} = y^i + e(\pi(i))$, $i = 1, \ldots, n$. Furthermore, a triangulation $\mathcal{T}$ is *symmetric* if $\sigma \in \mathcal{T}$ implies $-\sigma \in \mathcal{T}$. An example of symmetric integral triangulations of $\mathbb{R}^n$ is the K'-triangulation of Todd [31].

Now we introduce the class of simplicially local gross direction-preserving functions on $\mathbb{Z}^n$ on which the existence theorems of this paper are based. Locally gross direction-preservingness replaces the continuity condition for the existence of a zero point of a function defined on $\mathbb{R}^n$. Let $a \cdot b$ denote the inner product of two n-vectors a and b.

DEFINITION 2.1. (i) *A function $f : \mathbb{Z}^n \to \mathbb{R}^n$ is* locally gross direction-preserving *if, for any cell-connected points x and y in $\mathbb{Z}^n$,*

$$f(x) \cdot f(y) \geq 0.$$

(ii) *A function $f : \mathbb{Z}^n \to \mathbb{R}^n$ is* simplicially local gross direction-preserving *with respect to some given integral triangulation $\mathcal{T}$ of $\mathbb{R}^n$ if, for any vertices x and y of a simplex of $\mathcal{T}$,*

$$f(x) \cdot f(y) \geq 0.$$

The locally gross direction-preserving property was originally introduced in Yang [38] and prevents the function from changing too drastically in direction within one cell. The simplicially local gross direction-preserving condition is weaker and requires only that the function does not change too drastically in direction within any integral simplex of the given integral triangulation. Since any two vertices of a simplex of an integral triangulation are cell-connected, we have the property that every locally gross direction-preserving function is also simplicially local gross direction-preserving with respect to any integral triangulation.

To compute a discrete zero point of a simplicially local gross direction-preserving function, we adapt the $2n$-ray vector labeling algorithm of van der Laan and Talman [18] (see also Reiser [26] for integer labeling) to the current discrete setting. Let f be a simplicially local gross direction-preserving function with respect to some given integral triangulation $\mathcal{T}$ of $\mathbb{R}^n$. Let v be an arbitrarily chosen integral vector in $\mathbb{Z}^n$. The point v will be the starting point of the algorithm. For a nonzero sign vector $s \in \{-1, 0, +1\}^n$, the subset $A(s)$ of $\mathbb{R}^n$ is defined by

$$A(s) = \left\{ x \in \mathbb{R}^n \mid x = v + \sum_{h \in N} \alpha_h s_h e(h),\ \alpha_h \geq 0,\ h \in N \right\}.$$

Clearly, the set $A(s)$ is a t-dimensional subset of $\mathbb{R}^n$, where t is the number of nonzero components of the sign vector s, i.e., $t = |\{i \mid s_i \neq 0\}|$. Since $\mathcal{T}$ is an integral triangulation of $\mathbb{R}^n$, it triangulates every set $A(s)$ into t-dimensional integral simplices. For some s with t nonzero components, denote $\{h_1, \ldots, h_{n-t}\} = \{h \mid s_h = 0\}$, and let $\sigma = \langle x^1, \ldots, x^{t+1} \rangle$ be a t-simplex of the triangulation in $A(s)$. Following Todd [32], who improved the original system of equations used by van der Laan and Talman [18], we say that σ is *almost s-complete* if there is an $(n+2) \times (n+1)$ matrix W satisfying

$$\begin{bmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 & 0 \\ f(x^1) & \cdots & f(x^{t+1}) & e(h_1) & \cdots & e(h_{n-t}) & -s \end{bmatrix} W = I \tag{2.1}$$

and having rows $w^1, \ldots, w^{n+2}$ such that $w^h \succeq 0^{n+1}$ for $1 \leq h \leq t+1$, $w^{n+2} \succeq w^i$ and $w^{n+2} \succeq -w^i$ for $t+1 < i \leq n+1$, and $w^{n+2} \succeq 0^{n+1}$. Here I denotes the identity matrix of rank $n+1$. If $w_1^{n+2} = 0$, then we say that the simplex σ is complete. Further, let τ be a facet of σ, and without loss of generality, index the vertices of σ such that $\tau = \langle x^1, \ldots, x^t \rangle$. We say that τ is *s-complete* if there is an $(n+1) \times (n+1)$ matrix W satisfying

$$\begin{bmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 & 0 \\ f(x^1) & \cdots & f(x^t) & e(h_1) & \cdots & e(h_{n-t}) & -s \end{bmatrix} W = I \tag{2.2}$$

and having rows $w^1, \ldots, w^{n+1}$ such that $w^h \succeq 0^{n+1}$ for $1 \leq h \leq t$, $w^{n+1} \succeq w^i$ and $w^{n+1} \succeq -w^i$ for $t+1 \leq i \leq n$, and $w^{n+1} \succeq 0^{n+1}$. If $w_1^{n+1} = 0$, then we say that τ is complete.

The lemma below says that the zero-dimensional simplex $\langle v \rangle$ is an s^0-complete facet for a uniquely determined sign vector s^0. Let $\alpha = \max_h |f_h(v)|$. If $f_h(v) = -\alpha$ for some h, then we take $s_k^0 = -1$, where k is the smallest index h such that $f_h(v) = -\alpha$, and $s_j^0 = 0$ for $j \neq k$. If $f_h(v) > -\alpha$ for all h, then we take $s_k^0 = 1$, where k is the largest index h such that $f_h(v) = \alpha$, and $s_j^0 = 0$ for $j \neq k$. Let σ^0 be the unique one-dimensional simplex in $A(s^0)$ containing $\langle v \rangle$ as a facet. Clearly, s^0 contains only one nonzero element.

LEMMA 2.2. *The simplex $\langle v \rangle$ is an s^0-complete facet of σ^0. Moreover, s^0 is uniquely determined.*

Proof. Consider the system

$$\begin{bmatrix} 1 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 \\ f(v) & e(1) & \cdots & e(k-1) & e(k+1) & \cdots & e(n) & -s_k^0 e(k) \end{bmatrix} V = I.$$

Clearly, the first matrix on the left-hand side is regular, and therefore its inverse exists and equals the matrix V. The rows of V are given by

$$v^1 = (1, 0, \ldots, 0),$$

$$v^h = (-f_{h-1}(v), 0, \ldots, 0, 1, 0, \ldots, 0), \quad h = 2, \ldots, k \text{ if } k > 1,$$

with 1 being the hth component,

$$v^h = (-f_{h-1}(v), 0, \ldots, 0, 1, 0, \ldots, 0), \quad h = k+1, \ldots, n \text{ if } k < n,$$

with 1 being the $(h+1)$th component, and

$$v^{n+1} = (s_k^0 f_k(v), 0, \ldots, 0, -s_k^0, 0, \ldots, 0),$$

with $-s_k^0$ being the $(k+1)$th component. Clearly, v^1 is lexicographically positive. Moreover, v^{n+1} is lexicographically positive, because we have either $s_k^0 f_k(v) > 0$ or $s_k^0 f_k(v) = 0$ and $-s_k^0 > 0$. For $j = 2, \ldots, k$, we have $v^{n+1} \succeq v^j$, because $s_k^0 f_k(v) > 0$ and $s_k^0 f_k(v) > -f_{j-1}(v)$, and we also have $v^{n+1} \succeq -v^j$, because $s_k^0 f_k(v) > 0$ and $s_k^0 f_k(v) > f_{j-1}(v)$. For $j = k+1, \ldots, n$ and $s_k^0 = -1$, we have $v^{n+1} \succeq v^j$, because either $s_k^0 f_k(v) > 0$ and $s_k^0 f_k(v) > -f_j(v)$ or $s_k^0 f_k(v) = -f_j(v)$ and the $(j+1)$th component of v^j is 0, but the same component of v^{n+1} is 1; we also have $v^{n+1} \succeq -v^j$, because either $s_k^0 f_k(v) > 0$ and $s_k^0 f_k(v) > f_j(v)$ or $s_k^0 f_k(v) = f_j(v)$ and the $(j+1)$th component of v^j is 0, but the same component of v^{n+1} is 1. For $j = k+1, \ldots, n$ and $s_k^0 = 1$, we have $v^{n+1} \succeq v^j$, because $s_k^0 f_k(v) > 0$ and $s_k^0 f_k(v) > -f_j(v)$, and we also have $v^{n+1} \succeq -v^j$, because either $s_k^0 f_k(v) > 0$ and $s_k^0 f_k(v) > f_j(v)$ or $s_k^0 f_k(v) = f_j(v)$ and the $(j+1)$th component of $-v^j$ is -1, but the same component of v^{n+1} is 0. Hence, V satisfies all the requirements of the matrix W in system (2.2), and thus $\langle v \rangle$ is an s^0-complete facet of σ^0. Clearly, there is no other sign-vector s for which $\langle v \rangle$ is s-complete. □

We are now able to describe the algorithm for finding an integral solution to the system of equations $f(x) = 0^n$. When for some nonzero sign vector s a t-simplex $\sigma = \langle x^1, \ldots, x^{t+1} \rangle$ in $A(s)$ is almost s-complete, the system (2.1) has two "basic solutions." At each of these solutions exactly one row of the solution matrix W is binding. If $w_1^{n+2} = 0$, then σ is complete. If $w^h \succeq 0^{n+1}$ is binding for some h, $1 \le h \le t+1$, then the facet τ of σ opposite the vertex x^h is s-complete, and so τ is either (i) the zero-dimensional simplex $\langle v \rangle$ or (ii) a facet of precisely one other almost s-complete t-simplex σ' of the triangulation in $A(s)$ or (iii) τ lies on the boundary of $A(s)$ and is an almost s'-complete $(t-1)$-simplex in $A(s')$ for some unique nonzero sign vector s' with $t-1$ nonzero elements differing from s in only one element. If $w^{n+2} \succeq w^i$ ($w^{n+2} \succeq -w^i$) is binding for some $t+1 < i \le n+1$, σ is an s'-complete facet of precisely one almost s'-complete $(t+1)$-simplex in $A(s')$ for some nonzero sign vector s' differing from s in only the ith element, namely, $s'_i = +1$ (-1).

Since $\langle v \rangle$ is s^0-complete, σ^0 is an almost s^0-complete one-dimensional simplex in $A(s^0)$. Starting with σ^0, the $2n$-ray algorithm generates a sequence of adjacent almost s-complete simplices in $A(s)$ with s-complete common facets for varying sign vectors s. Moving from one s-complete facet to the next s'-complete facet corresponds to making a lexicographic linear programming pivot step from one of the two basic solutions of system (2.1) to the other. The algorithm stops as soon as it finds a complete simplex. We will show that in that case one of its vertices is a discrete zero point of the function f.

LEMMA 2.3. *Let f be simplicially local gross direction-preserving with respect to $\mathcal{T}$. Then any complete simplex contains a discrete zero point of the function f.*

Proof. Let $x^1, \ldots, x^{k+1}$ be the vertices of a complete simplex σ in $A(s)$, and let t be the number of nonzeros in s. Notice that $k = t-1$ or $k = t$ depending on whether σ is a t-simplex in $A(s)$ or a facet of a t-simplex in $A(s)$. From the system (2.1) or (2.2) it follows that there exists $\lambda_1 \geq 0, \ldots, \lambda_{k+1} \geq 0$ with sum equal to 1 such that $\sum_{j=1}^{k+1} \lambda_j f(x^j) = 0^n$. Let j^* be such that $\lambda_{j^*} > 0$. Then by premultiplying $f(x^{j^*})$ on both sides of $\sum_{j=1}^{k+1} \lambda_j f(x^j) = 0^n$, we obtain

$$\lambda_1 f(x^1) \cdot f(x^{j^*}) + \cdots + \lambda_{j^*} f(x^{j^*}) \cdot f(x^{j^*}) + \cdots + \lambda_{k+1} f(x^{k+1}) \cdot f(x^{j^*}) = 0.$$

Since f is simplicially local gross direction-preserving, it is easy to see that every term in the above expression is nonnegative. Therefore every term is equal to 0. In particular, $f(x^{j^*}) = 0^n$, and so x^{j^*} is a discrete zero point of the function f. ∎

Formally, the steps of the above algorithm are given below in detail.

1. Initial Step: Compute $f(v)$. If $f(v) = 0^n$, then the algorithm terminates with v as a solution. Otherwise $\langle v \rangle$ is an s^0-complete facet of a unique 1-simplex $\sigma^0 = \langle v, v^+ \rangle$ in $A(s^0)$. Let $s = s^0$, $t = |\{i \mid s_i \neq 0\}|$, and $\sigma = \sigma^0$. Go to main step 1 with the system (2.1) corresponding to σ^0.
2. Main Step 1: Perform a lexicographic linear programming (LLP) pivot step in the system (2.1) with the column $(1, f(v^+))$. If $w_1^{n+2} = 0$, the algorithm terminates with a complete simplex which yields a solution. Otherwise, in the case that $w^h \succeq 0^{n+1}$ is binding for some h, $1 \leq h \leq t+1$, then the facet τ of σ opposite the vertex x^h is s-complete, and go to main step 2. In the case that $w^{n+2} \succeq w^i$ ($w^{n+2} \succeq -w^i$) is binding for some $t+1 < i \leq n+1$, go to main step 3.
3. Main Step 2: If τ is a facet of precisely one other almost s-complete t-simplex σ' of the triangulation in $A(s)$, let v^+ be the vertex of σ' differing from those of τ, let $\sigma = \sigma'$, and go to main step 1. Otherwise, τ lies on the boundary of $A(s)$ and is an almost s'-complete $(t-1)$-simplex in $A(s')$ for some unique nonzero sign vector s' with $t-1$ nonzero elements differing from s in only one element. Let h be the unique element with $s_h \neq 0$ and $s'_h = 0$, $\sigma = \tau$ and $s = s'$, and go to main step 4.
4. Main Step 3: σ is an s'-complete facet of precisely one almost s'-complete $(t+1)$-simplex σ' in $A(s')$ for some nonzero sign vector s' differing from s in only the ith element, namely, $s'_i = +1$ (-1). Let v^+ be the vertex of σ' differing from those of σ, $s = s'$ and $\sigma = \sigma'$, and go to main step 1.
5. Main Step 4: Perform an LLP pivot step in the system (2.1) with the column $(0, e(h))$. If $w_1^{n+2} = 0$, the algorithm terminates with a complete simplex which yields a solution. Otherwise, in the case that $w^h \succeq 0^{n+1}$ is binding for some h, $1 \leq h \leq t+1$, then the facet τ of σ opposite the vertex x^h is s-complete, and go to main step 2. In the case that $w^{n+2} \succeq w^i$ ($w^{n+2} \succeq -w^i$) is binding for some $t+1 < i \leq n+1$, go to main step 3.

Because all steps are uniquely determined due to the lexicographically pivoting and the properties of a triangulation, the algorithm cannot visit any simplex more than once, and therefore either the algorithm terminates in a finite number of iterations with a complete simplex yielding a solution or the sequence of simplices generated by the algorithm goes to infinity. In the next section we present a convergence condition which prevents the latter case from happening and thus ensures the existence of a solution.

3. Convergence conditions. To present a convergence condition for the algorithm, for $x \in \mathbb{Z}^n$, let $N(x)$ denote the set of integer points being simplicially connected to x.

Assumption 3.1 (convergence condition). Given a function $f : \mathbb{Z}^n \to \mathbb{R}^n$, there exist vectors m, $M \in \mathbb{Z}^n$, with $m_h < M_h - 1$ for every $h \in N$, such that for every integral vector x on the boundary of the set $C^n = \{z \in \mathbb{R}^n \mid m \leq z \leq M\}$ the following conditions hold:

(i) If $x_i = m_i$, then $f_i(y) \geq 0$ for all $y \in N(x) \cap C^n$ satisfying $y_i = m_i$, or there exists $j \in N$ such that $f_j(y) < f_i(y)$ for all $y \in N(x) \cap C^n$ satisfying $y_i = m_i$.

(ii) If $x_i = M_i$, then $f_i(y) \leq 0$ for all $y \in N(x) \cap C^n$ satisfying $y_i = M_i$, or there exists $j \in N$ such that $f_j(y) > f_i(y)$ for all $y \in N(x) \cap C^n$ satisfying $y_i = M_i$.

The condition means that there exist lower and upper bounds such that, when x is an integral vector on the ith lower (upper) bound, then either $f_i(y)$ is nonnegative (nonpositive) for any integral vector y on the same lower (upper) bound being simplicially connected to x or, for some $j \neq i$, $f_i(y)$ is bigger (smaller) than $f_j(y)$ for any integral vector y on the same lower (upper) bound being simplicially connected to x. We show that under this condition any simplicially local gross direction-preserving function has a discrete zero point within the bounded set C^n induced by the lower and upper bounds. To do so, the starting point v of the $2n$-ray algorithm is taken to be an arbitrarily chosen integral vector in the interior of the set C^n. Then the constructive proof of Theorem 3.2 is based on the combinatorial argument that under the convergence condition the algorithm cannot cross the boundary of the set C^n, and therefore it must terminate in a finite number of steps with a simplex having one of its vertices as an integral solution to f. It is worth pointing out that while both the lower and upper bounds are part of the condition in the theorem, in our constructive proof we need only the starting point to lie between these bounds without need to know exactly what they are. Typically, in applications these bounds are naturally determined and indicate the domain of interest underlying the problem; see, for instance, the application in section 6.

THEOREM 3.2. *Let $f : \mathbb{Z}^n \to \mathbb{R}^n$ be a simplicially local gross direction-preserving function with respect to some integral triangulation $\mathcal{T}$. If f satisfies Assumption* 3.1*, then f has a discrete zero point.*

Proof. Take any integral vector in the interior of the set C^n as the starting point v of the algorithm. By definition of integral triangulation, $\mathcal{T}$ triangulates the set C^n and also the set $A(s) \cap C^n$ for any sign vector s into integral simplices.

For some nonzero sign vector s, let τ be an s-complete facet in $A(s)$ with vertices $x^1, \ldots, x^t$, where t is the number of nonzeros in s. We first show that τ is complete if it is on the boundary of C^n. From system (2.2) it follows that there exist $\lambda_1 \geq 0, \ldots, \lambda_t \geq 0$ with sum equal to 1, $\beta \geq 0$, and $-\beta \leq \mu_i \leq \beta$ for $s_i = 0$ such that $\bar{f}_i(z) = \beta$ if $s_i = 1$, $\bar{f}_i(z) = -\beta$ if $s_i = -1$, and $\bar{f}_i(z) = \mu_i$ if $s_i = 0$, where $z = \sum_{i=1}^t \lambda_i x^i$ and $\bar{f}(z) = \sum_{i=1}^t \lambda_i f(x^i)$; i.e., $\bar{f}$ is the piecewise linear extension of f with respect to $\mathcal{T}$. Since τ lies on the boundary of C^n, there exists an index h such that either $x_h^j = m_h$ for all j or $x_h^j = M_h$ for all j. In case $x_h^j = m_h$ for all j, we have $s_h = -1$ and therefore $\bar{f}_h(z) = -\beta$. Furthermore, by Assumption 3.1, we have (i) $f_h(x^j) \geq 0$ for all j or (ii) there exists k such that $f_k(x^j) < f_h(x^j)$ for all j. In case (ii) we obtain $\bar{f}_k(z) < \bar{f}_h(z)$. On the other hand, $\bar{f}_k(z) \geq -\beta = \bar{f}_h(z)$, yielding a contradiction; i.e., this case cannot occur. In case (i) we obtain $\bar{f}_h(z) \geq 0$. On the other hand $\bar{f}_h(z) = -\beta \leq 0$. Therefore $\bar{f}_h(z) = 0$ and also $\beta = 0$. Since $w_1^{n+1} = \beta$ we obtain that τ is complete. Similarly, we can show that the same results hold for the case of $x_h^j = M_h$ for all j.

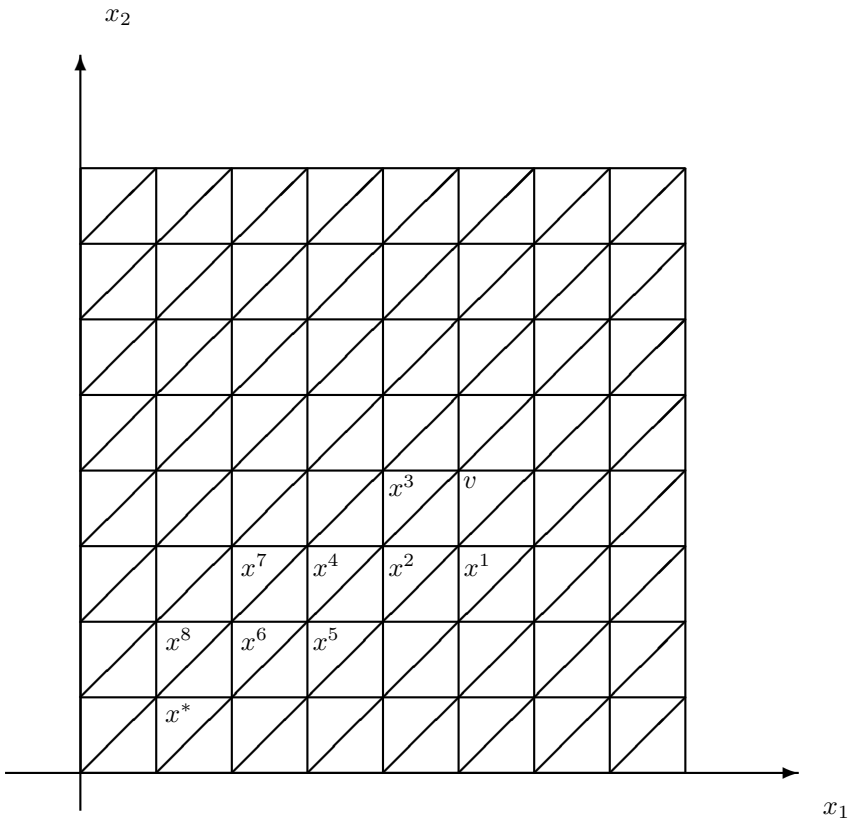


FIG. 1. *Illustration of the algorithm.*

Now, consider the algorithm as described at the end of the previous section. Due to the lexicographic pivoting rule, the algorithm will never visit any simplex more than once. So, because the number of simplices in C^n is finite, the algorithm finds in a finite number of steps a complete simplex. Since f is simplicially local gross direction-preserving, Lemma 2.3 guarantees that at least one of the vertices of this simplex is a discrete zero point of the function f. □

We conclude this section with an example to illustrate the conditions of the theorem and how the algorithm operates. Consider the function $f : \mathbb{Z}^2 \to \mathbb{R}^2$ defined by $f(x) = (2 - 2x_1, x_1 - x_2^2)$. This function is simplicially locally gross direction-preserving with respect to the K-triangulation described in the previous section. It is interesting to note that f is not locally gross direction-preserving since, e.g., for the cell-connected points $x = (1, 2)$ and $y = (2, 1)$, we have $f(x) = (0, -3)$ and $f(y) = (-2, 1)$ and so $f(x) \cdot f(y) = -3 < 0$. Further, the example satisfies Assumption 3.1 for any vector $m = (a, a)$ and $M = (b, b)$ with $a < 0$ and $b > 1$, implying that the convergence condition of Theorem 3.2 is satisfied. Hence, there exists a solution, and in fact $x^* = (1, 1)$ is a discrete zero point. Let the starting point v be $(5, 4)$. Then the sequence of points traced by the algorithm is shown in Figure 1 and given by $x^1 = (5, 3)$, $x^2 = (4, 3)$, $x^3 = (4, 4)$, $x^4 = (3, 3)$, $x^5 = (3, 2)$, $x^6 = (2, 2)$, $x^7 = (2, 3)$, and $x^8 = (1, 2)$ and leads to the solution $x^* = (1, 1)$ in 10 function evaluations. Observe that to apply the algorithm we do not need to fix the bounds a priori.

The following corollary strengthens a result of Yang [38] for locally gross direction-preserving functions and follows immediately from Theorem 3.2.

COROLLARY 3.3. *Let $f : \mathbb{Z}^n \to \mathbb{R}^n$ be a simplicially local gross direction-preserving function. Suppose that there exist vectors m, $M \in \mathbb{Z}^n$, with $m_h < M_h - 1$ for every $h \in N$, such that for every integral vector x on the boundary of the set $C^n = \{z \in \mathbb{R}^n \mid m \le z \le M\}$, $x_i = m_i$ implies $f_i(x) \ge 0$ and $x_i = M_i$ implies $f_i(x) \le 0$. Then f has a discrete zero point.*

Furthermore, we have the following discrete fixed point theorem.

COROLLARY 3.4. *Let $D^n = \{z \in \mathbb{Z}^n \mid m \le z \le M\}$, where m and M are vectors in $\mathbb{Z}^n$ with $m_h < M_h - 1$ for every $h \in N$. Assume that $f : D^n \to \mathrm{Co}(D^n)$ is a function such that $x - f(x)$ is a simplicially local gross direction-preserving function in x. Then f has a discrete fixed point.*

Proof. Define the function $g : D^n \to \mathbb{R}^n$ by $g(x) = x - f(x)$. Clearly, g satisfies the condition of Corollary 3.3. So there exists $x^* \in D^n$ such that $g(x^*) = 0$, i.e., $f(x^*) = x^*$. □

4. Convergency under an antipodal condition. In this section we modify the algorithm in section 2 to find a discrete zero point under a general antipodal condition to be stated next.

Assumption 4.1 (antipodal condition). Given a function $f : \mathbb{Z}^n \to \mathbb{R}^n$, there exists a vector $u \in \mathbb{Z}^n$ with $u_h \ge 1$ for all $h \in N$ such that $f(x) \cdot f(-y) \le 0$ for any cell-connected integral points x and y lying on a same proper face of the set $U^n = \{z \in \mathbb{R}^n \mid -u \le z \le u\}$.

This condition is very natural and might be viewed as a discrete analogue of a weak version of the Borsuk–Ulam antipodal condition for a continuous function saying that $f(x) \cdot f(-x) \le 0$ when x is on the boundary of U^n. It is known that under the latter condition a continuous function has a zero point; see for instance van der Laan [16] and Yang [36]. Todd and Wright [33] used a modification of the $2n$-ray algorithm to give a constructive proof of the Borsuk–Ulam theorem and Freund and Todd [9] used the modified algorithm to give a constructive proof for a combinatorial lemma of Tucker [34]. Yang [38] proposed the antipodal condition and showed that under the condition a locally gross direction-preserving function has a discrete zero point. The next theorem strengthens this result by allowing for simplicially local gross direction-preservingness on a symmetric triangulation in the interior of U^n.

THEOREM 4.2. *Let $f : \mathbb{Z}^n \to \mathbb{R}^n$ be a simplicially local gross direction-preserving function with respect to a symmetric integral triangulation $\mathcal{T}$ of $\mathbb{R}^n$, satisfying Assumption* 4.1 *and $f(x) \cdot f(y) \ge 0$ for any cell-connected integral points x and y lying on a same proper face of the set U^n. Then f has a discrete zero point.*

The next example illustrates the theorem. Let $f : \mathbb{Z}^2 \to \mathbb{R}^2$ be given by $f(x) = (x_1 - x_2 - 1,\ x_2 - 1)$. Then it is easy to see that f satisfies the antipodal property for $u_1 = u_2 = 4$. Further, it is easy to check that f is simplicially local gross direction-preserving with respect to the symmetric integral K'-triangulation of $\mathbb{R}^2$ (see section 2) on the interior of U^n and f is locally gross direction-preserving on the boundary of U^n, as required in the last condition of the theorem. So, f has a discrete zero point and in fact the point $(2, 1)$ is the unique discrete zero point. Observe that f is not locally gross direction-preserving in the interior of U^n and therefore the existence does not follow from the result of Yang [38]. For instance, for the cell-connected points $x = (2, 0)$ and $y = (1, 1)$, we have $f(x) \cdot f(y) = -1 < 0$.

Besides the relaxation to simplicially local gross direction-preserving, the main contribution of this section is that, in contrast to the nonconstructive proof in [38], below we give a constructive proof for the theorem. We now modify the $2n$-ray algorithm of section 2 to accomodate the antipodal condition. The modification is

based on a lemma on the extension V^n of the set U^n given by

$$V^n = \{x \in \mathbb{R}^n \mid -(u_i + 1) \le x_i \le u_i + 1 \ \ \forall i \in N\}.$$

Let the projection function $p : V^n \to U^n$ be defined by

$$p_h(x) = \max\{-u_h, \min\{u_h, x_h\}\} \ \ \forall h \in N.$$

Clearly, $p(x) = x$ if $x \in U^n$. Moreover, $p(x) \in U^n \cap \mathbb{Z}^n$ if $x \in V^n \cap \mathbb{Z}^n$. We now extend f to the function $g : V^n \cap \mathbb{Z}^n \to \mathbb{R}^n$ by setting $g(x) = f(x)$ for $x \in U^n$ and $g(x) = f(p(x)) - f(-p(x))$ for $x \in V^n \setminus U^n$. It follows straightforwardly that $g(x) = -g(-x)$ for any $x \in \mathbb{Z}^n \cap \mathrm{Bd}(V^n)$. We now have the following lemma.

LEMMA 4.3. *For $f : \mathbb{Z}^n \to \mathbb{R}^n$ as given in Theorem* 4.2*, the extension g of f to V^n is simplicially local gross direction-preserving on $\mathbb{Z}^n \cap V^n$ with respect to the given symmetric triangulation $\mathcal{T}$.*

Proof. Clearly, g is simplicially local gross direction-preserving on U^n. It remains to consider the following two cases.

First, let $x, y \in \mathbb{Z}^n$ be two vertices of a simplex of $\mathcal{T}$ on the boundary of V^n. Then $p(x)$ and $p(y)$ are two cell-connected points on a same proper face of U^n and thus satisfy $f(p(x)) \cdot f(p(y)) \ge 0$. The same holds for $-p(x)$ and $-p(y)$. Together with the antipodal Assumption 4.1 this yields

$$\begin{aligned} g(x) \cdot g(y) &= (f(p(x)) - f(-p(x))) \cdot (f(p(y)) - f(-p(y))) \\ &= f(p(x)) \cdot f(p(y)) - f(p(x)) \cdot f(-p(y)) - f(-p(x)) \cdot f(p(y)) \\ &\quad + f(-p(x)) \cdot f(-p(y)) \ge 0. \end{aligned}$$

Second, let $x, y \in \mathbb{Z}^n$ be two vertices of a simplex of $\mathcal{T}$ with x on the boundary of U^n and y on the boundary of V^n. Again x and $p(y)$ are two cell-connected points on a same proper face of U^n and thus $f(x) \cdot f(p(y)) \ge 0$. Together with the antipodal condition this again yields

$$\begin{aligned} g(x) \cdot g(y) &= f(x) \cdot (f(p(y)) - f(-p(y))) \\ &= f(x) \cdot f(p(y)) - f(x) \cdot f(-p(y)) \ge 0. \end{aligned}$$ □

Proof of Theorem 4.2. To prove the theorem, let the set V^n and the function g be defined as above. Take the origin 0^n of $\mathbb{R}^n$ as the starting point v of the algorithm as described in section 2. The underlying symmetric integral triangulation $\mathcal{T}$ for the function f subdivides each set $A(s)$ into t-simplices such that if σ is a simplex in $A(s)$, then $-\sigma$ is a simplex in $A(-s)$.

Starting with the origin, the algorithm generates a sequence of adjacent almost s-complete simplices with s-complete common facets in $A(s) \cap V^n$ for varying sign vectors s with the following modification. If in the main step 2 of the algorithm τ is an s-complete facet lying in $A(s)$ on the boundary of V^n, then the antipodal facet $-\tau$ is a $(-s)$-complete facet in $A(-s)$ on the boundary of V^n, since $g(x) = -g(-x)$ for any $x \in \mathbb{Z}^n \cap \mathrm{Bd}(V^n)$. The algorithm continues with main step 1 by letting $s = -s$, σ the unique almost $(-s)$-complete simplex in $A(-s) \cap V^n$ containing $-\tau$ as a facet and v^+ the vertex of σ opposite to facet $-\tau$. The algorithm therefore always stays in V^n and due to the lexicographic pivoting rule will never visit any simplex in V^n more than once. Since the number of simplices in V^n is finite, within a finite number of steps the algorithm terminates with a complete simplex σ^* in V^n. Since g is simplicially local gross direction-preserving, by the Lemmas 2.3 and 4.3 it follows that σ^* has a vertex z being a discrete zero point of g. It remains to prove that $p(z) \in U^n$ is a discrete

zero point of f. If z is not on the boundary of V^n, then $p(z) = z$ is an integral vector in U^n and $g(z) = f(z)$, and therefore z is a discrete zero point of f. Suppose z is on the boundary of V^n. Since $g(z) = 0^n$, this implies

$$\begin{aligned} 0 &= f(p(z)) \cdot g(z) = f(p(z)) \cdot (f(p(z)) - f(-p(z))) \\ &= f(p(z)) \cdot f(p(z)) - f(p(z)) \cdot f(-p(z)), \end{aligned}$$

and therefore

$$0 \leq f(p(z)) \cdot f(p(z)) = f(p(z)) \cdot f(-p(z)) \leq 0,$$

where the last inequality follows from the antipodal condition on f. Hence, $f(p(z)) \cdot f(p(z)) = 0$ and therefore $p(z)$ is a discrete zero point of f in U^n. □

5. A method for discrete complementarity problems. The complementarity problem is to find a point $x^* \in \mathbb{R}^n$ such that

$$x^* \geq 0^n, \quad f(x^*) \geq 0^n, \quad \text{and} \quad x^* \cdot f(x^*) = 0,$$

where f is a given function from $\mathbb{R}^n$ into itself. For an arbitrary function f, the problem is called the *nonlinear complementarity problem*. In case f is affine, i.e., $f(x) = Mx + q$ with M being an $n \times n$ matrix and q being an n-vector, the problem is called the *linear complementarity problem*, denoted by LCP(M, q). There is by now a voluminous literature on the complementarity problem; see Lemke [22], Cottle [2], Karamardian [13], Moré [24] and [25], Kojima [14], and van der Laan and Talman [19], among others. For comprehensive surveys on the subject, see, e.g., Kojima et al. [15], Cottle, Pang, and Stone [3], and Facchinei and Pang [7].

In the following we consider the problem that the solution of the complementarity problem is required to be integral or that the function f is defined only on the integer lattice $\mathbb{Z}^n$ of $\mathbb{R}^n$. In this case we call the problem the *discrete complementarity problem*, denoted by DCP(f). We first give sufficient conditions under which the general case DCP(f) has a solution, and we will give a constructive proof of this existence result by modifying the system of equations of the algorithm in section 2 to the current situation. Next we will show that when applied to the linear complementarity problem LCP(M, q), the algorithm reduces to the well-known Lemke's method [22] and finds an integral solution provided that M is totally unimodular and copositive-plus, and the system of $Mx + q \geq 0^n$, $x \geq 0^n$, is feasible.

In the following, for any $x, y \in \mathbb{R}^n$, let $I(x) = \{i \mid x_i > 0\}$, and let $I(x, y) = I(x) \cup I(y)$. We first modify the definition of simplicially local gross direction-preservingness for points on the boundary of the nonnegative orthant $\mathbb{R}^n_+$.

DEFINITION 5.1. *A function $f : \mathbb{Z}^n \to \mathbb{R}^n$ is* simplicially local gross direction-preserving *with respect to some given integral triangulation $\mathcal{T}$ of $\mathbb{R}^n$ if, for any two vertices x and y of a simplex of $\mathcal{T}$ in $\mathbb{R}^n_+$, it holds that*

$$f_i(x) f_i(y) \geq 0 \ \textit{whenever} \ x_i = y_i = 0 \ \textit{and} \ \sum_{h \in I(x,y)} f_h(x) f_h(y) \geq 0.$$

The next theorem establishes the existence of a solution to DCP(f) under a natural condition.

THEOREM 5.2. *Let $f : \mathbb{Z}^n \to \mathbb{R}^n$ be a simplicially local gross direction-preserving function on $\mathbb{Z}^n_+$. If there exists a vector $M \in \mathbb{Z}^n_{++}$ such that, for any $x \in \mathbb{Z}^n_+$ with $x \leq M$, $x_i = M_i$ implies $f_i(x) \geq 0$, then DCP(f) has a solution.*

We will provide a constructive proof by applying the algorithm in section 2 to the current situation. To do so, the origin 0^n is taken as the starting point v. Since 0^n lies on the boundary of $\mathbb{R}^n_+$, the sets $A(s)$ and s-completeness are defined only for nonnegative nonzero sign vectors s. Notice that $A(s) = \{x \in \mathbb{R}^n_+ \mid x_i = 0 \text{ whenever } s_i = 0\}$. Further, to apply the algorithm in this case, we have to adapt the concepts of an almost s-complete simplex and an s-complete facet. For some sign vector s with t positive components, denote $\{h_1, \ldots, h_{n-t}\} = \{h \mid s_h = 0\}$, and let $\sigma = \langle x^1, \ldots, x^{t+1}\rangle$ be a t-simplex of the triangulation in $A(s)$. Then σ is *almost s-complete* if there is an $(n+2) \times (n+1)$ matrix W being a solution to system

$$\begin{bmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 & 0 \\ f(x^1) & \cdots & f(x^{t+1}) & -e(h_1) & \cdots & -e(h_{n-t}) & s \end{bmatrix} W = I \tag{5.1}$$

and having rows $w^1, \ldots, w^{n+2}$ such that $w^h \succeq 0^{n+1}$ for $1 \le h \le t+1$, $w^{n+2} \succeq -w^i$ for $t+1 < i \le n+1$, and $w^{n+2} \succeq 0^{n+1}$. If $w^{n+2}_1 = 0$, then we say that the simplex σ is complete. For τ a facet of σ, without loss of generality, letting $\tau = \langle x^1, \ldots, x^t\rangle$, τ is *s-complete* if there is an $(n+1) \times (n+1)$ matrix W being a solution to system

$$\begin{bmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 & 0 \\ f(x^1) & \cdots & f(x^t) & -e(h_1) & \cdots & -e(h_{n-t}) & s \end{bmatrix} W = I \tag{5.2}$$

and having rows $w^1, \ldots, w^{n+1}$ such that $w^h \succeq 0^{n+1}$ for $1 \le h \le t$, $w^{n+1} \succeq -w^i$ for $t+1 \le i \le n$, $w^{n+1} \succeq 0^{n+1}$. If $w^{n+1}_1 = 0$, then we say that τ is complete.

With respect to the starting point 0^n, let $\alpha = \min_h f_h(0^n)$, and let s^0 be the sign vector with $s^0_k = 1$, where k is the smallest index h such that $f_h(0^n) = \alpha$, and $s^0_j = 0$ for $j \ne k$. To avoid triviality, we may assume that $f(0^n) \not\succeq 0^n$. Similarly as in section 2, it can be shown that the simplex $\langle 0^n\rangle$ is an s^0-complete facet of the unique one-dimensional simplex σ^0 in $A(s^0)$ having $\langle 0^n\rangle$ as one of its facets. Furthermore σ^0 is almost s^0-complete.

We now apply the algorithm as described in section 2. Starting with σ^0, by applying the steps as given in section 2 with the system (5.1) the algorithm generates a unique sequence of adjacent almost s-complete simplices in $A(s)$ with s-complete common facets for varying positive sign vectors s. The algorithm stops with a complete simplex in a finite number of steps under the assumption stated in the theorem. As shown below, a complete simplex gives a solution to the problem. Recall that $\bar{f}$ stands for the piecewise linear extension of the function f with respect to $\mathcal{T}$. Let $C^n = \{x \in \mathbb{R}^n \mid 0^n \le x \le M\}$.

LEMMA 5.3. *For some nonnegative sign vector s, let σ be a simplex in $A(s)$ with an s-complete facet τ on the upper boundary of C^n. Then τ is a complete simplex.*

Proof. From system (5.2) it follows that there exist $\lambda_1 \ge 0, \ldots, \lambda_t \ge 0$ with sum equal to 1, $\beta \ge 0$, and $\mu_i \ge -\beta$ for $s_i = 0$, such that $\bar{f}_i(z) = -\beta$ when $s_i = 1$ and $\bar{f}_i(z) = \mu_i$ when $s_i = 0$, where $z = \sum_{i=1}^t \lambda_i x^i$. Since τ lies on the upper boundary of C^n, there exists an index h such that $x^j_h = M_h$ for all j. But then we must have $s_h = 1$ and therefore $\bar{f}_h(z) = -\beta \le 0$. On the other hand, by assumption, we have $f_h(x^j) \ge 0$ for all j. Hence, we obtain $\bar{f}_h(z) \ge 0$. As a result, $\beta = 0$, which implies that τ is a complete simplex by definition. ∎

LEMMA 5.4. *For some nonnegative sign vector s, let σ be a complete simplex in $A(s)$. Then σ contains a solution to the nonlinear complementarity problem for $\bar{f}$.*

Proof. Let $x^1, \ldots, x^{k+1}$ be the vertices of the complete simplex σ in $A(s)$, and let t be the number of nonzeros in s. Note that $k = t-1$ or $k = t$ depending on whether

σ is a t-simplex in $A(s)$ or a facet of a t-simplex in $A(s)$. It follows from the system (5.1) or (5.2) that there exist $\lambda_1 \geq 0, \ldots, \lambda_{k+1} \geq 0$ with sum equal to 1 and $\mu_i \geq 0$ for $s_i = 0$ such that $\bar{f}_i(z) = 0$ if $s_i = 1$, and $\bar{f}_i(z) = \mu_i$ if $s_i = 0$, where $z = \sum_{i=1}^{k+1} \lambda_i x^i$. Since $z \in A(s)$, we also have $z_i = 0$ if $s_i = 0$ and $z_i \geq 0$ if $s_i = 1$. So, $\bar{f}_i(z) \geq 0$ if $z_i = 0$ and $\bar{f}_i(z) = 0$ if $z_i > 0$; i.e., z solves the nonlinear complementarity problem with respect to $\bar{f}$. □

The next lemma says that, for any complete simplex σ, at least one of its vertices is a solution to DCP(f).

LEMMA 5.5. *Let σ be a complete simplex of $\mathcal{T}$ in $A(s)$ for some sign vector s. Then σ contains a vertex being a solution to* DCP(f).

Proof. Because σ is a complete simplex in $A(s)$, as shown in Lemma 5.4, there is a point z in σ that is a solution to the nonlinear complementarity problem with respect to $\bar{f}$. Now let $\rho = \langle x^1, \ldots, x^k \rangle$ be the unique face of σ containing z in its relative interior. Namely, there exist unique positive numbers $\lambda_1, \ldots, \lambda_k$ summing up to 1 such that $z = \sum_{j=1}^{k} \lambda_j x^j$ and $\bar{f}(z) = \sum_{j=1}^{k} \lambda_j f(x^j)$. Take any j^* between 1 and k. Suppose first that $z_i = 0$ and $\bar{f}_i(z) > 0$ for some i. Clearly, $x_i^j = 0$ for all $j = 1, \ldots, k$. Since $\bar{f}_i(z) = \sum_{j=1}^{k} \lambda_j f_i(x^j)$ there exists h such that $f_i(x^h) > 0$. Since x^h and x^{j^*} are simplicially connected and $x_i^h = x_i^{j^*} = 0$, we have that $f_i(x^h) f_i(x^{j^*}) \geq 0$, and therefore $x_i^{j^*} = 0$ and $f_i(x^{j^*}) \geq 0$. Suppose next that $z_i = 0$ and $\bar{f}_i(z) = 0$ for some i. Again, $x_i^j = 0$ for all $j = 1, \ldots, k$. Since $\bar{f}_i(z) = \sum_{j=1}^{k} \lambda_j f_i(x^j)$ and $\bar{f}_i(z) = 0$, we obtain $\sum_{j=1}^{k} \lambda_j f_i(x^j) = 0$ and therefore $\sum_{j=1}^{k} \lambda_j f_i(x^j) f_i(x^{j^*}) = 0$. Since for all j it holds that x^j and x^{j^*} are simplicially connected and $x_i^j = x_i^{j^*} = 0$, we have $f_i(x^j) f_i(x^{j^*}) \geq 0$, and so each term in the summation must be zero. In particular, it holds that $\lambda_{j^*} f_i^2(x^{j^*}) = 0$. Since $\lambda_{j^*} > 0$, this implies $f_i(x^{j^*}) = 0$.

Thus far we have shown that whenever $z_i = 0$ both $f_i(x^{j^*}) \geq 0$ and $x_i^{j^*} = 0$ must hold. It remains to show that whenever $z_i > 0$ it holds that $f_i(x^{j^*}) = 0$ and hence that x^{j^*} is a solution to DCP(f). By construction, $\bar{f}_i(z) = \sum_{j=1}^{k} \lambda_j f_i(x^j) = 0$ whenever $z_i > 0$. Note that $I(x^j) \subseteq I(z)$ for any $j = 1, \ldots, k$. Therefore,

$$\sum_{h \in I(z)} \sum_{j=1}^{k} \lambda_j f_h(x^j) f_h(x^{j^*}) = 0$$

and so

$$\sum_{j=1}^{k} \left(\lambda_j \sum_{h \in I(z)} f_h(x^j) f_h(x^{j^*}) \right) = 0.$$

Since $I(z)$ contains the set $I(x^j, x^{j^*})$ and x^j and x^{j^*} are simplicially connected for all j, by hypothesis we have that each of the k terms between brackets is nonnegative and therefore must be zero. Hence,

$$\lambda_{j^*} \sum_{h \in I(z)} f_h^2(x^{j^*}) = 0.$$

Since $\lambda_{j^*} > 0$, we obtain $f_h(x^{j^*}) = 0$ for all $h \in I(z)$. Therefore $f_i(x^{j^*}) = 0$ if $z_i > 0$. □

Theorem 5.2 now follows from the lemmas stated above by a combinatorial argument.

Proof of Theorem 5.2. Due to the lexicographic pivoting rule, the algorithm will never visit any simplex more than once. Since the number of simplices in C^n is finite, the algorithm terminates in a finite number of steps with a complete simplex in $A(s)$. According to Lemma 5.5, the complete simplex gives a solution to DCP(f). □

In what follows, we turn our attention to the linear complementarity problem LCP(M, q). Recall that Lemke's method [22] introduces an artificial variable z_0 and operates by moving from one basic solution of the following system of linear equations to another:

$$\begin{aligned} Iz - Mx - z_0 e &= q, \\ x_j \geq 0,\ z_j \geq 0,\ z_0 \geq 0 &\text{ for } j = 1, 2, \ldots, n, \\ x_j z_j &= 0 \text{ for } j = 1, 2, \ldots, n, \end{aligned} \tag{5.3}$$

where e is the n-vector of all ones and I is the identity matrix of order n. The algorithm starts with a ray at $x = 0^n$ and terminates in a finite number of pivot steps when a solution is found or when another ray is encountered.

Lemke [22] shows that his method is guaranteed to find a solution of LCP(M, q) if M is copositive-plus and the system of $Mx + q \geq 0^n$ and $x \geq 0^n$ is feasible. Recall that a square matrix B is said to be *copositive* if $x \cdot Bx \geq 0$ for any $x \in \mathbb{R}^n_+$. Furthermore, B is said to be *copositive-plus* if B is copositive and in addition $x \geq 0^n$ and $x \cdot Bx = 0$ imply $(B + B^t)x = 0$, where B^t is the transpose of B. Of course, even if an LCP(M, q) has a solution, it may have no integral solution at all. To guarantee that an LCP(M, q) has an integral solution, we need to impose total unimodularity on the matrix M. Recall that a matrix B is said to be *totally unimodular* if the determinant of every subsquare matrix of B is -1, 0, or 1. Now we establish the following theorem on the existence of an integral solution to LCP(M, q).

THEOREM 5.6. *Suppose that M is totally unimodular, copositive-plus, that q is an integral vector, and that the system of $Mx + q \geq 0^n$ and $x \geq 0^n$ is feasible. Then the algorithm defined by system* (5.1) *reduces to Lemke's method and terminates at an integral solution in a finite number of steps.*

Proof. For the LCP(M, q), we first show that the algorithm defined by system (5.1) reduces to Lemke's method. We may assume that $q \not\geq 0^n$. In the initial step of Lemke's method the system defined by (5.3) at $x = 0^n$ is put in a tableau format, and a pivot step is made with the z_0 column on row k, where k is such that $q_k = \min\{q_h \mid h \in N\}$. This corresponds exactly to the initial step of the algorithm defined by system (5.1) at which 0^n is the starting point and the algorithm moves in the set $A(s^0)$, where $s^0 \in \mathbb{R}^n_+$ is the sign vector defined by $s^0_k = 1$ and $s^0_j = 0$ for $j \neq k$ with k being the smallest index h such that $q_h = \min q_j$. Here the choice of the smallest index is to avoid degeneracy.

In a general step, let $\sigma = \langle x^1, \ldots, x^{t+1} \rangle$ be an almost s-complete simplex in $A(s)$. Let $I^0(s) = \{h \mid s_h = 0\}$ and $I^+(s) = \{h \mid s_h = 1\}$. Now it follows from the system (5.1) that there exist $\lambda_h \geq 0$, $h = 1, \ldots, t+1$, $\mu_0 \geq 0$, and $\mu_0 \geq -\mu_h$ for every $h \in I^0(s)$ such that

$$\sum_j \lambda_j f(x^j) - \sum_{h \in I^0(s)} \mu_h e(h) + \mu_0 s = 0^n \tag{5.4}$$

and $\sum_j \lambda_j = 1$. Let $\beta_0 = \mu_0$ and $\beta_h = -\mu_h$. Let $x = \sum_j \lambda_j x^j$ for $h \in I^0(s)$. Since x is a convex combination of points $x^1, \ldots, x^{t+1}$ in $A(s)$, x also lies in $A(s)$. Further,

$f(x) = Mx + q = \sum_j \lambda_j (Mx^j + q) = \sum_j \lambda_j f(x^j)$. Thus equation (5.4) reduces to

$$Mx + q + \sum_{h \in I^0(s)} \beta_h e(h) + \beta_0 s = 0^n,$$

where $\beta_0 \geq 0$, $\beta_0 \geq \beta_h$ for every $h \in I^0(s)$, $x_h = 0$ for $s_h = 0$, and $x_h \geq 0$ for $s_h = 1$. We can rewrite the equation as follows:

$$-Mx + \sum_{h \in I^0(s)} (\beta_0 - \beta_h) e(h) - \beta_0 e = q,$$

Now let $z_h = \beta_0 - \beta_h$ for $h \in I^0(s)$, $z_h = 0$ for $h \in I^+(s)$, and $z_0 = \beta_0$. Then we have

$$\begin{aligned} Iz - Mx - z_0 e &= q, \\ z_0 \geq 0,\ z_h &\geq 0 \text{ for } h \in I^0(s), \\ z_h &= 0 \text{ for } h \in I^+(s), \\ x_h &= 0 \text{ for } h \in I^0(s), \\ x_h &\geq 0 \text{ for } h \in I^+(s). \end{aligned} \tag{5.5}$$

For this system we have $x_h z_h = 0$ for every $h = 1, \ldots, n$, and the algorithm finds a solution as soon as z_0 becomes zero. This shows that the system above coincides with the system (5.3). As a result, we have proved that the algorithm defined by system (5.1) indeed reduces to Lemke's method. It is worth pointing out that for the LCP(M, q), actually no triangulation is needed for the algorithm. In fact, for given sign vector s, all pivot steps of the $2n$-ray algorithm within the region $A(s)$ reduce to one pivot step in the Lemke algorithm because of the linearity of the function $f(x) = Mx + q$.

Concerning the second statement of the Theorem 5.6 (the termination of the algorithm), it follows from Lemke [22] that because M is copositive-plus and the system of $Mx + q \geq 0^n$ and $x \geq 0^n$ is feasible, the algorithm must end up with a solution x^* in a finite number of steps. More precisely, the algorithm stops with a solution $x^* \in A(s)$ for some sign vector $s \in \mathbb{R}^n_+$ corresponding to the $n \times n$ regular matrix $B = [(-M_h, h \in I^+(s)), (e(h), h \in I^0(s))]$, where M_h denotes the hth column of matrix M. Note that $x^* = B^{-1}q$. It remains to show that x^* is integral. Because M is totally unimodular, $[-M, I]$ is totally unimodular and so is B (see Schrijver [29]). Because B^{-1} exists and is also totally unimodular and q is integral, $x^* = B^{-1}q$ is integral. This shows that the algorithm indeed finds an integral solution of the LCP(M, q). □

Since a positive definite matrix M is copositive-plus and it holds that there exists $x \geq 0^n$ such that $Mx + q \geq 0^n$ (see Cottle, Pang, and Stone [3, p. 140, Lemma 3.1.3]), the next corollary follows immediately.

COROLLARY 5.7. *Suppose that M is totally unimodular, positive definite, and that q is an integral vector. Then the algorithm terminates at an integral solution in a finite number of steps.*

6. Applications. Discrete zero point (or fixed point) problems often occur in economics. For instance, in an exchange economy with n commodities, one of the most studied problems is the existence of a Walrasian equilibrium price system, in which a price vector $p \in \mathbb{R}^n_+$ solve the complementarity problem for the excess demand function z defined from the price space $\mathbb{R}^n_+$ into the commodity space $\mathbb{R}^n$, where $z_j(p)$ is the excess demand for commodity j at the (nonnegative) price system p, $j = 1, \ldots, n$. In

the literature the existence of an equilibrium price system $p^* \in \mathbb{R}^n_+$ has been studied extensively; nevertheless, in almost all real-life situations prices are in some monetary unit, implying that actually a price system belongs to $\mathbb{Z}^n_+$ for appropriately chosen units of the components of p. Hence, in fact an equilibrium price system should be a solution to the DCP(z).

In this section we apply Theorem 3.2 to the supermodular games; see for instance Fudenberg and Tirole [10]. A well-known example of such games is the Bertrand price competition model. Here we consider the Cournot oligopoly model with complementary commodities; see Vives [35]. There are n firms, each firm producing its own commodity. The goal of each firm is to choose an amount of product that maximizes its own profit given the production levels chosen by other firms. Let $q_i \geq 0$ denote the quantity of commodity i produced by firm i, and let $q_{-i} = (q_1, \ldots, q_{i-1}, q_{i+1}, \ldots, q_n)$ denote the vector of amounts of commodities produced by all firms but firm i. The price at which firm i can sell its product is decreasing in its own quantity q_i, and due to the complementarities, it is increasing in the quantities q_j, $j \neq i$. It is standard to assume that the price function of each firm $i = 1, \ldots, n$, is linear, i.e.,

$$P_i(q_i, q_{-i}) = a_i - b_i q_i + \sum_{j \neq i} d_{ij} q_j,$$

where all parameters a_i, b_i, and d_{ij} are positive. Each firm i has a linear cost function $C_i(q_i) = c_i q_i$ with $a_i > c_i > 0$. For quantities $(q_1, \ldots, q_n)$, the profit π_i of firm i is given by its quantity times price minus its cost of production, i.e.,

$$\pi_i(q_i, q_{-i}) = q_i P_i(q_i, q_{-i}) - c_i q_i.$$

A tuple $(q_1^*, q_2^*, \ldots, q_n^*) \in \mathbb{R}^n_+$ of nonnegative real numbers is a *Cournot–Nash equilibrium* if, for every firm i,

$$\pi_i(q_i^*, q_{-i}^*) \geq \pi_i(q_i, q_{-i}^*) \ \forall \ q_i \in \mathbb{R}_+.$$

It is well known that there exists a Cournot–Nash equilibrium if $2b_i > \sum_{j \neq i} \ d_{ij}$ for every firm $i = 1, \ldots, n$. However, in reality, it is often the case that the commodities are produced only in integer quantities. Here we will show that under the same condition a discrete Cournot–Nash equilibrium exists in this model. A tuple $(q_1^*, q_2^*, \ldots, q_n^*)$ of nonnegative integers is a *discrete Cournot–Nash equilibrium* if

$$\pi_i(q_i^*, q_{-i}^*) \geq \pi_i(q_i, q_{-i}^*) \ \forall \ q_i \in \mathbb{Z}_+, \ i = 1, \ldots, n.$$

That is, given the quantities chosen by other firms, each firm chooses an integer quantity that yields a profit which is at least as high as any other integer quantity could give.

For a real number x, the symbol $[x]$ denotes the greatest nearest integer to x. Given nonnegative integer quantities q_{-i} of all other firms, firm i maximizes its own profit $\pi_i(q_i, q_{-i})$ over all nonnegative integers q_i, and its optimal or reaction integer quantity is given by

$$r_i(q_{-i}) = \left[\frac{a_i - c_i}{2b_i} + \sum_{j \neq i} \frac{d_{ij}}{2b_i} q_j \right].$$

Observe that $r_i(q_{-i}) \geq 0$ for all $q \in \mathbb{Z}^n_+$, because $a_i > c_i > 0$. Define the function $f : \mathbb{Z}^n_+ \to \mathbb{Z}^n$ by

$$f_i(q_i, q_{-i}) = r_i(q_{-i}) - q_i, \quad i = 1, \ldots, n.$$

Clearly, a discrete zero point of f is a discrete Cournot–Nash equilibrium.

THEOREM 6.1. *Suppose that $2b_i > \sum_{j \neq i} d_{ij}$, $i = 1, \ldots, n$. Then there exists a discrete Cournot–Nash equilibrium in the above Cournot oligopoly competition model.*

Proof. We show that the function f satisfies the conditions of Corollary 3.3. First, we show that f satisfies the boundary condition. As a natural lower bound, take $m = 0^n$, and as an upper bound take, for all i, $M_i = M$, $i = 1, \ldots, n$, with $M > 1$ an integer satisfying $M > \max_i\{(a_i - c_i)/(2b_i - \sum_{j \neq i} d_{ij})\}$. Then for any i and any $q \in \mathbb{Z}^n_+$, $q_i = 0$ implies $f_i(q) = r_i(q_{-i}) \geq 0$. Further, $q_i = M$ and $q_j \leq M$ imply

$$\begin{aligned} f_i(q) &= \left[\frac{a_i - c_i}{2b_i} + \sum_{j \neq i} \frac{d_{ij}}{2b_i} q_j - q_i \right] \leq \left[\frac{a_i - c_i}{2b_i} + \sum_{j \neq i} \frac{d_{ij}}{2b_i} M - M \right] \\ &\leq \left[\frac{a_i - c_i - (2b_i - \sum_{j \neq i} d_{ij})M}{2b_i} \right] \leq 0, \end{aligned}$$

where the last inequality follows from the fact that $M > (a_i - c_i)/(2b_i - \sum_{j \neq i} d_{ij})$.

Second, we show that f is simplicially local gross direction-preserving with respect to the K-triangulation as described in section 2. Since the K-triangulation is given by integral simplices $\sigma(y, \pi)$ with vertices $y^1, \ldots, y^{n+1}$, with $y^1 = y$ and $y^{i+1} = y^i + e(\pi(i))$, $i = 1, \ldots, n$, for given $y \in \mathbb{Z}^n$ and $\pi = (\pi(1), \ldots, \pi(n))$ a permutation of the elements $1, 2, \ldots, n$, we have to check that $f(x) \cdot f(y) \geq 0$ for any pair $x \in \mathbb{Z}^n_+$ and $y = x + \sum_{h=1}^k e(\pi(h))$ for $k = 1, \ldots, n$ and any permutation π. Observe that for any such pair it holds that $y_i \in \{x_i, x_i + 1\}$ for all $i = 1, \ldots, n$. For some pair x, y and $i \in \{1, \ldots, n\}$, define $S_i(x, y) = \{j \neq i | y_j = x_j + 1\}$. Then $r_i(y) = [\frac{a_i - c_i}{2b_i} + \sum_{j \neq i} \frac{d_{ij}}{2b_i} y_j] = [\frac{a_i - c_i}{2b_i} + \sum_{j \neq i} \frac{d_{ij}}{2b_i} x_j + \sum_{j \in S_i(x,y)} \frac{d_{ij}}{2b_i}]$. Since $\sum_{j \neq i} \frac{d_{ij}}{2b_i} < 1$, it follows that $r_i(y) \in \{r_i(x), r_i(x) + 1\}$. Hence, since $y_i \in \{x_i, x_i + 1\}$, it follows that $f_i(y) \in \{f_i(x) - 1, f_i(x), f_i(x) + 1\}$. So, when $f_i(x) \geq 1$, then $f_i(y) \geq f_i(x) - 1 \geq 0$, and when $f_i(x) \leq -1$, then $f_i(y) \leq f_i(x) + 1 \leq 0$. So, $f_i(x) f_i(y) \geq 0$ for all i and thus $f(x) \cdot f(y) \geq 0$.

We have shown that f satisfies all the conditions of Corollary 3.3 and thus has a discrete zero point. As a result, there is a discrete Cournot–Nash equilibrium. ∎

It is worth mentioning that f may not be simplicially local gross direction-preserving with respect to other triangulations, as shown by the following example with $n = 2$. Let the parameters be given by $a_1 = 4$, $c_1 = 2.5$, $b_1 = 0.5$, $d_1 = d_{12} = 3/4$, $a_2 = 5$, $c_2 = 4$, $b_2 = 1/3$, and $d_2 = d_{21} = 1/12$. These parameters satisfy the stated condition for the model, and therefore there is a discrete Cournot–Nash equilibrium. In fact, the quantities $(3, 2)$ form the unique discrete Cournot–Nash equilibrium for this example. As shown above, the function f is simplicially local gross direction-preserving with respect to the K-triangulation. However, this function is not simplicially local gross direction-preserving with respect to the H-triangulation of Saigal [27]. For $\mathbb{R}^2$, this triangulation is given by the simplices $\langle y^1, y^2, y^3 \rangle$, with $y^1 \in \mathbb{Z}^2$, $y^2 = y^1 + p(\pi(1))$, $y^3 = y^2 + p(\pi(2))$, where $p(1) = (1, 0)$ and $p(2) = (-1, 1)$. Now take $\pi = (2, 1)$, $x = y^1 = (3, 1)$, and $y = y^2 = y^1 + p(2) = (2, 2)$. Since $f(x) = (-1, 1)$ and $f(y) = (1, 0)$, we have that $f(x) \cdot f(y) = -1 < 0$, and so the function is not

simplicially local direction-preserving with respect to the H-triangulation. Note that x and y do not belong to a same simplex of the K-triangulation.

Acknowledgments. We would like to thank the referees for several helpful suggestions. This research was carried out while Gerard van der Laan and Zaifu Yang were visiting the CentER for Economic Research, Tilburg University. The visit of Zaifu Yang has been made possible by financial support of CentER and the Netherlands Organization for Scientific Research (NWO).

REFERENCES

[1] E. L. Allgower and K. Georg, *Numerical Continuation Methods: An Introduction*, Springer, Berlin, 1990.
[2] R. W. Cottle, *Nonlinear programs with positively bounded Jacobians*, SIAM J. Appl. Math., 14 (1966), pp. 147–158.
[3] R. W. Cottle, J.-S. Pang, and R. E. Stone, *The Linear Complementarity Problem*, Academic Press, New York, 1992.
[4] V. Danilov and G. Koshevoy, *Existence Theorem of Zero Point in a Discrete Case*, manuscript, 2004, pp. 1–5.
[5] B. C. Eaves, *Homotopies for computation of fixed points*, Math. Program., 3 (1972), pp. 1–22.
[6] B. C. Eaves and R. Saigal, *Homotopies for computation of fixed points on unbounded regions*, Math. Program., 3 (1972), pp. 225–237.
[7] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Vols. I and II, Springer, New York, 2003.
[8] H. Freudenthal, *Simplizialzerlegungen von beschrankter flachheit*, Ann. Math., 43 (1942), pp. 580–582.
[9] R. M. Freund and M. J. Todd, *A constructive proof of Tucker's combinatorial lemma*, J. Combin. Theory Ser. A, 30 (1981), pp. 321–325.
[10] D. Fudenberg and J. Tirole, *Game Theory*, MIT Press, Boston, 1993.
[11] T. Iimura, *A discrete fixed point theorem and its applications*, J. Math. Econom., 39 (2003), pp. 725–742.
[12] T. Iimura, K. Murota, and A. Tamura, *Discrete fixed point theorem reconsidered*, J. Math. Econom., 41 (2005), pp. 1030–1036.
[13] S. Karamardian, *The complementarity problem*, Math. Program., 2 (1972), pp. 107–129.
[14] M. Kojima, *A unification of the existence theorems of the nonlinear complementarity problem*, Math. Program., 9 (1975), pp. 257–277.
[15] M. Kojima, N. Megiddo, T. Noma, and A. Yoshise, *A Unified Approach to Interior Point Algorithms for Linear Complementarity Problems*, Springer, Berlin, 1991.
[16] G. van der Laan, *On the existence and approximation of zeros*, Math. Program., 28 (1984), pp. 1–14.
[17] G. van der Laan and A. J. J. Talman, *A restart algorithm for computing fixed points without an extra dimension*, Math. Program., 17 (1979), pp. 74–84.
[18] G. van der Laan and A. J. J. Talman, *A class of simplicial restart fixed point algorithms without an extra dimension*, Math. Program., 20 (1981), pp. 33–48.
[19] G. van der Laan and A. J. J. Talman, *Simplicial approximation of solutions to the nonlinear complementarity problem*, Math. Program., 38 (1987), pp. 1–15.
[20] G. van der Laan, A. J. J. Talman, and Z. Yang, *Solving discrete zero point problems*, Math. Program., 108 (2006), pp. 127–134.
[21] G. van der Laan, A. J. J. Talman, and Z. Yang, *Computing integral solutions of complementarity problems*, discussion paper TI 2005-006/1, Tinbergen Institute, Amsterdam, 2005.
[22] C. E. Lemke, *Bimatrix equilibrium points and mathematical programming*, Management Sci., 11 (1965), pp. 681–689.
[23] O. H. Merrill, *Applications and Extensions of an Algorithm that Computes Fixed Points of Certain Upper Semi-Continuous Point-to-Set Mappings*, Ph.D. thesis, University of Michigan, Ann Arbor, 1972.
[24] J. J. Moré, *Coercivity conditions in nonlinear complementarity problem*, SIAM Rev., 17 (1974), pp. 1–16.
[25] J. J. Moré, *Classes of functions and feasibility conditions in nonlinear complementarity problem*, Math. Program., 6 (1974), pp. 327–338.

[26] P. M. Reiser, *A modified integer labeling for complementarity algorithms*, Math. Oper. Res., 6 (1981), pp. 129–139.
[27] R. Saigal, *Investigations into the efficiency of fixed point algorithms*, in Fixed Points: Algorithms and Applications, S. Karamardian, ed., Academic Press, New York, 1977, pp. 203–223.
[28] H. Scarf, *The approximation of fixed points of a continuous mapping*, SIAM J. Appl. Math., 15 (1967), pp. 1328–1343.
[29] A. Schrijver, *Theory of Linear and Integer Programming*, John Wiley and Sons, Chichester, 1986.
[30] M. J. Todd, *Computation of Fixed Points and Applications*, Springer, Berlin, 1976.
[31] M. J. Todd, *Improving the convergence of fixed point algorithms*, Math. Program. Stud., 7 (1978), pp. 151–179.
[32] M. J. Todd, *Global and local convergence and monotonicity results for a recent variable-dimension simplicial algorithm*, in Numerical Solution of Highly Nonlinear Problems, W. Forster, ed., North–Holland, Amsterdam, 1980.
[33] M. J. Todd and A. H. Wright, *A variable dimension simplicial algorithm for antipodal fixed point theorems*, Numer. Funct. Anal. Optim., 2 (1980), pp. 155–186.
[34] A. W. Tucker, *Some topological properties of disk and sphere*, in Proceedings of the First Canadian Mathematical Congress (Montreal, 1945), University of Toronto Press, Toronto, 1946, pp. 285–309.
[35] X. Vives, *Complementarities and games: new developments*, J. Econom. Lit., 43 (2005), pp. 437-479.
[36] Z. Yang, *Computing Equilibria and Fixed Points*, Kluwer, Boston, 1999.
[37] Z. Yang, *Discrete Nonlinear Complementarity Problems*, Math. Oper. Res., to appear.
[38] Z. Yang, *Discrete Fixed Point Analysis and Its Applications*, FBA working paper 210, Yokohama National University, Yokohama, 2004.

SIAM J. OPTIM.
Vol. 18, No. 1, pp. 309–321

CHARACTERIZATIONS OF LOCAL AND GLOBAL ERROR BOUNDS FOR CONVEX INEQUALITIES IN BANACH SPACES*

HUI HU†

Abstract. This paper studies local and global error bounds for a convex inequality defined by a proper convex function in a Banach space. The concept of weak basic constraint qualification (weak BCQ) is introduced to control the normal directions at a boundary point of the solution set. Local and global error bounds are characterized by a direction-length decomposition condition, which provides a way to independently verify the weak BCQ and the length control of the subdifferential. To further characterize global error bounds, the segment extension property is proposed and studied. It is shown that the verification of the direction-length condition for global error bounds can be sufficiently carried out on any subset having the segment extension property instead of the entire boundary. This leads to a simple formula for the smallest global error bound. In the Euclidean space, the verification of the condition and the computation of the smallest global error bound can be carried out on the set of extreme points.

Key words. local and global error bounds, weak basic constraint qualification, segment extension property, end set

AMS subject classifications. 90C, 90C25

DOI. 10.1137/050644872

1. Introduction.

Let X be a Banach space, let R be the set of real numbers, and let $f : X \to R \cup \{\infty\}$ be a proper convex function. For the convex inequality

$$f(x) \le 0, \tag{1.1}$$

let $S = \{x \in X : f(x) \le 0\}$ denote the solution set and $\mathrm{bd}(S)$ the boundary of S. Throughout this paper, it is assumed that S is closed and $\emptyset \ne S \ne X$. The study of a closed convex set defined by a non–lower-semicontinuous function arose from a broad class of outer approximation methods for convex optimization (see [3] and the references therein).

For any $P \subseteq X$ and $x \in X$, let $d_P(x) = \inf\{\|x - p\| : p \in P\}$ if $P \ne \emptyset$; otherwise, $d_P(x) = \infty$ by convention. Let $f_+(x) = \max\{f(x), 0\}$ and $R_+ = \{r \in R : r \ge 0\}$.

Let $x \in \mathrm{bd}(S)$. Inequality (1.1) is said to have a local error bound τ [8, 16, 19] at x if there exist $\tau, \delta \in (0, \infty)$ such that

$$d_S(z) \le \tau f_+(z) \quad \forall z \in B(x, \delta),$$

where $B(x, \delta)$ denotes the open ball centered at x with radius δ.

Inequality (1.1) is said to have a global error bound τ [12] if there exists $\tau \in (0, \infty)$ such that

$$d_S(z) \le \tau f_+(z) \quad \forall z \in X.$$

In recent years, local and global error bounds and their relations with other key concepts in optimization have been actively studied by many researchers (see, e.g., [7, 8, 9, 10, 12, 16, 19] and the references therein).

*Received by the editors November 10, 2005; accepted for publication (in revised form) October 10, 2006; published electronically April 17, 2007.
http://www.siam.org/journals/siopt/18-1/64487.html
†Department of Mathematical Sciences, Northern Illinois University, DeKalb, IL 60115 (hu@math.niu.edu).

This paper studies local and global error bounds for a convex inequality defined by a single-valued proper convex function in a Banach space. In section 2, some useful results for dealing with discontinuity on $\mathrm{bd}(S)$ and nonclosed epigraphs are presented. In section 3, the concept of weak basic constraint qualification (weak BCQ) is introduced to control the normal directions at a boundary point of S without lower semicontinuity assumption on f. Local and global error bounds are characterized by a general direction-length decomposition condition, which provides a way to independently verify the direction control by the weak BCQ and the length control of the intersection of the normal cone and the end set of the subdifferential. This decomposition condition can reduce the verification difficulty. In addition, the length control provides a way to compute the error bounds. In section 4, the segment extension property is proposed and studied. It is shown that the verification of the direction-length condition for global error bounds can be sufficiently carried out on any subset having the segment extension property instead of the entire boundary. This leads to a simple formula for computing the smallest global error bound. In the Euclidean space, the verification of the condition and the computation of the smallest global error bound can be carried out on the set of extreme points.

The notation used in this paper is standard. For $P \subseteq X$, let $\overline{P}$, $\mathrm{conv}(P)$, $\mathrm{cone}(P)$, $\mathrm{int}(P)$ denote the closure, convex hull, convex cone, and interior of P. If P is convex, let $\mathrm{ext}(P)$ denote the set of extreme points of P, and P^∞ the recession cone of P.

For a convex set C in X and $x \in C$, let

$$N_C(x) = \{x^* \in X^* : \ \langle x^*, y - x\rangle \le 0 \ \ \forall y \in C\}$$

denote the normal cone of C at x, where X^* is the dual space of X and $\langle x^*, y - x\rangle$ denotes the value of linear functional x^* at $y - x$.

For $x \in \mathrm{bd}(S)$, let

$$T_S(x) = \overline{\cup_{t>0}\, t(S - x)}$$

denote the tangent cone of S at x.

Let $\mathrm{dom} f = \{x \in X : \ f(x) < \infty\}$ and $\mathrm{epi} f = \{(x, y) \in X \times R : \ f(x) \le y\}$. The subdifferential and singular subdifferential of f at $x \in \mathrm{dom} f$ are (cf. [5, 19])

$$\partial f(x) = \{x^* \in X^* : \ (x^*, -1) \in N_{\mathrm{epi} f}(x, f(x))\},$$

$$\partial^\infty f(x) = \{x^* \in X^* : (x^*, 0) \in N_{\mathrm{epi} f}(x, f(x))\}.$$

It is easy to verify that

$$\partial f(x) = \{x^* \in X^* : f(y) \ge f(x) + \langle x^*, y - x\rangle \ \forall y \in X\},$$

$$\partial^\infty f(x) = N_{\mathrm{dom} f}(x), \tag{1.2}$$

and if $\partial f(x) \ne \emptyset$, then

$$\partial f(x) = \partial f(x) + \lambda \partial^\infty f(x) \quad \forall \lambda \in [0, \infty). \tag{1.3}$$

Let $f'(x; h)$ denote the classical directional derivative of f at x in the direction h, i.e.,

$$f'(x; h) = \lim_{t \to 0^+} \frac{f(x + th) - f(x)}{t}.$$

For a nonempty interval $[\alpha,\beta]$ and a convex set C, if $C\neq\emptyset$, then $[\alpha,\beta]C=\{tc: t\in[\alpha,\beta],\ c\in C\}$; otherwise, $[\alpha,\beta]C=\{0\}$ by convention. It is easy to verify that $[0,1]C=\operatorname{conv}(\{0\}\cup C)$.

For a convex set C, the end set of C is defined as

$$E[C]=\{z\in\overline{[0,1]C}:\ tz\notin\overline{[0,1]C}\ \forall t>1\}.$$

Note that $0\notin E[C]$ and $E[C]\cap\operatorname{int}(C)=\emptyset$. The end set was first introduced in [19] and was further studied in [6]. A referee of the present paper pointed out that the definition of the end set is equivalent to the following.

LEMMA 1.1. *For a convex set C, let $\tilde{E}[C]=\{z\in\overline{C}:\ tz\notin\overline{C}$ for all $t>1\}$. Then $E[C]=\tilde{E}[C]$.*

Proof. First, we show that $E[C]\subseteq\tilde{E}[C]$. If $z\in E[C]$, then $z\in\overline{[0,1]C}$ and $tz\notin\overline{[0,1]C}$ for all $t>1$. In particular, $tz\notin\overline{C}$ for all $t>1$. It remains to show that $z\in\overline{C}$. There exist sequences $\tau_n\in[0,1]$ and $c_n\in C$ such that $\tau_nc_n\to z$. By passing to a subsequence if necessary, one may assume that $\tau_n\to\tau\in[0,1]$. By the convexity of $\overline{[0,1]C}$, $\tau_nc_n+(1-\tau_n)z\in\overline{[0,1]C}$. Thus, $\lim_{n\to\infty}\tau_nc_n+(1-\tau_n)z=(2-\tau)z\in\overline{[0,1]C}$. By the definition of $E[C]$, $2-\tau\leq1$. Therefore, one mast have $\tau=1$. Consequently, $z=(\lim_{n\to\infty}\frac{1}{\tau_n})(\lim_{n\to\infty}\tau_nc_n)=\lim_{n\to\infty}c_n\in\overline{C}$.

Next, we show that $\tilde{E}[C]\subseteq E[C]$. Let $z\in\tilde{E}[C]$. Since $z\in\overline{C}\subseteq\overline{[0,1]C}$, we only need to show that $tz\notin\overline{[0,1]C}$ for all $t>1$. Suppose, on the contrary, that there exists $t^*>1$ such that $t^*z\in\overline{[0,1]C}$. Then there exist sequences $\tau_n\in[0,1]$ and $c_n\in C$ satisfying $\tau_nc_n\to t^*z$. By passing to a subsequence if necessary, one may assume that $\tau_n\to\tau\in[0,1]$. By the convexity of $\overline{C}$, $\tau_nc_n+(1-\tau_n)z\in\overline{C}$. Thus, $\lim_{n\to\infty}\tau_nc_n+(1-\tau_n)z=t^*z+(1-\tau)z=(t^*+1-\tau)z\in\overline{C}$. But $t^*+1-\tau>1$, which is a contradiction to $tz\notin\overline{C}$ for all $t>1$. Therefore, $\tilde{E}[C]\subseteq E[C]$. ∎

2. Analysis without lower semicontinuity. Note that (1.1) is only defined by a proper convex function. Without a lower semicontinuity assumption, $\operatorname{epi}f$ is not necessarily closed; therefore, some standard results cannot be applied directly. Also, it is possible that $f(x)<0$ for some $x\in\operatorname{bd}(S)$. In this section, we present several useful results for dealing with such cases. First, we state some facts that are frequently used in the rest of this paper. Throughout this paper, f is a proper convex function on X.

FACT 2.1. *For each $h\in X$, the quotient $\frac{f(x+th)-f(x)}{t}$ is a nondecreasing function of $t>0$. Thus, $f'(x;h)$ exists ($+\infty$ and $-\infty$ being allowed as limits) and*

$$f'(x;h)=\inf\Big\{\frac{f(x+th)-f(x)}{t}:t>0\Big\}\leq f(x+h)-f(x).$$

FACT 2.2. (i) *$x^*\partial f(x)\Leftrightarrow\langle x^*,h\rangle\leq f'(x;h)$ for all $h\in X$.* (ii) *If f is continuous at $x\in\operatorname{dom}f$, then $f'(x;h)=\max\{\langle x^*,h\rangle:x^*\partial f(x)\}$ for all $h\in X$ (cf.* [18, Theorem 2.4.9]*).*

FACT 2.3. *If $x\in\operatorname{bd}(S)$, then $N_S(x)=N_{x+T_S(x)}(x)=N_{T_S(x)}(0)$.*

FACT 2.4. *If $x\in\operatorname{dom}f$ and $\partial f(x)\neq\emptyset$, then (see, e.g.,* [19, pp. 761–762]*) $\partial^\infty f(x)\subseteq\overline{[0,1]\partial f(x)}=[0,1]\partial f(x)+\partial^\infty f(x)$.*

Note that if $f(x)=0$, then $\partial f(x)\subseteq N_S(x)$ holds. When (1.1) is defined by a proper convex function f that is not continuous, it is possible that $f(x)<0$ for $x\in\operatorname{bd}(S)$. In such a case, $\partial f(x)\subseteq N_S(x)$ may fail to hold (see Example 3.1). The following two lemmas are useful in dealing with such discontinuity.

LEMMA 2.1. *If $x \in \mathrm{bd}(S)$ and $f(x) < 0$, then* (i) $\mathrm{dom} f \subseteq x + T_S(x)$, (ii) $f'(x;h) = \infty$ *for all* $h \notin T_S(x)$.

Proof. Given $y \notin x + T_S(x)$, let $h = y - x \notin T_S(x)$. By the definition of tangent cone, $x + th \notin S$ for all $t > 0$, which implies that $f(x + th) > 0$ for all $t > 0$. By Fact 2.1,

$$f(y) - f(x) \geq f'(x;h) = \lim_{t \to 0^+} \frac{f(x+th) - f(x)}{t} \geq \lim_{t \to 0^+} \frac{-f(x)}{t} = \infty.$$

Therefore, $y \notin \mathrm{dom} f$ and $f'(x;h) = \infty$. □

LEMMA 2.2. *If $x \in \mathrm{bd}(S)$ and $f(x) < 0$, then*

$$N_S(x) = \partial^\infty f(x) = \partial^\infty f_+(x) = \partial f_+(x).$$

Proof. By Lemma 2.1, $S \subseteq \mathrm{dom} f = \mathrm{dom} f_+ \subseteq x + T_S(x)$. Hence,

$$N_S(x) \supseteq N_{\mathrm{dom} f}(x) = N_{\mathrm{dom} f_+}(x) \supseteq N_{x + T_S(x)}(x).$$

It follows from (1.2) and Fact 2.3 that

$$N_S(x) = \partial^\infty f(x) = \partial^\infty f_+(x). \tag{2.1}$$

Since $f_+(x) = 0$ and $f_+(z) \geq 0$ for all $z \in X$, x minimizes f_+ and thus $0 \in \partial f_+(x)$. By (1.3), $\partial^\infty f_+(x) \subseteq \partial f_+(x) \subseteq N_S(x)$. It follows from (2.1) that $\partial^\infty f_+(x) = \partial f_+(x) = N_S(x)$. □

Without lower semicontinuity, the epigraph of f may not be closed, and some standard results for closed convex sets cannot be applied directly. Lemma 2.3 summarizes two results for dealing with epigraphs that are not closed.

LEMMA 2.3. *Let $f_1 = f$ and $f_2 = 0$.*

(i) $\overline{\mathrm{epi} f_1} \cap \mathrm{epi} f_2 = \overline{\mathrm{epi} f_1 \cap \mathrm{epi} f_2}$.

(ii) *If $\mathrm{dom} f_1 \cap \mathrm{dom} f_2 \neq \emptyset$, then* $\mathrm{cone}(\mathrm{dom} f_1 - \mathrm{dom} f_2) \times R = \mathrm{cone}(\mathrm{epi} f_1 - \mathrm{epi} f_2)$.

Proof. (i) Let $z = (y, r) \in \overline{\mathrm{epi} f_1} \cap \mathrm{epi} f_2 \subseteq X \times R_+$, where $y \in X$ and $r \in R_+$. Choose $z_n = (y_n, r_n) \in \mathrm{epi} f_1$ satisfying $y_n \to y$ and $r_n \to r$. Let $\tilde{z}_n = (y_n, \max\{r_n, 0\})$. We have $\tilde{z}_n \in \mathrm{epi} f_1 \cap \mathrm{epi} f_2$ and $\tilde{z}_n \to (y, r) \in \overline{\mathrm{epi} f_1 \cap \mathrm{epi} f_2}$. On the other hand, since $\mathrm{epi} f_2 = X \times R_+$ is closed, $\overline{\mathrm{epi} f_1 \cap \mathrm{epi} f_2} \subseteq \overline{\mathrm{epi} f_2} = \mathrm{epi} f_2$. Thus, $\overline{\mathrm{epi} f_1 \cap \mathrm{epi} f_2} \subseteq \overline{\mathrm{epi} f_1} \cap \mathrm{epi} f_2$.

(ii) Given $\lambda((x_1, r_1) - (x_2, r_2)) \in \mathrm{cone}(\mathrm{epi} f_1 - \mathrm{epi} f_2)$, where $\lambda \geq 0$, $f(x_1) \leq r_1$, and $f(x_2) \leq r_2$, one has

$$\lambda((x_1, r_1) - (x_2, r_2)) = (\lambda(x_1 - x_2), \lambda(r_1 - r_2)) \in \mathrm{cone}(\mathrm{dom} f_1 - \mathrm{dom} f_2) \times R.$$

On the other hand, for a given $(\lambda(x_1 - x_2), r) \in \mathrm{cone}(\mathrm{dom} f_1 - \mathrm{dom} f_2) \times R$, where $\lambda \geq 0$, $x_1 \in \mathrm{dom} f_1$, and $x_2 \in \mathrm{dom} f_2$, we consider two cases: $\lambda > 0$ and $\lambda = 0$. If $\lambda > 0$, then

$$\begin{aligned}
(\lambda(x_1 - x_2), r) &= (\lambda x_1, 2r) - (\lambda x_2, r) \\
&= \lambda\big[(x_1, 2r/\lambda) - (x_2, r/\lambda)\big] \\
&= \lambda\big[(x_1, 2r/\lambda + M) - (x_2, r/\lambda + M)\big] \\
&\in \mathrm{cone}(\mathrm{epi} f_1 - \mathrm{epi} f_2),
\end{aligned}$$

where M is sufficiently large such that $(x_1, 2r/\lambda + M) \in \mathrm{epi} f_1$ and $(x_2, r/\lambda + M) \in \mathrm{epi} f_2$. In the case of $\lambda = 0$, for $x \in \mathrm{dom} f_1 \cap \mathrm{dom} f_2$, we have $(0, r) = (x, 2r) - (x, r) =$

$(x, 2r+M)-(x, r+M) \in \text{cone}(\text{epi}f_1 - \text{epi}f_2)$, where M is sufficiently large such that $(x, 2r+M) \in \text{epi}f_1$ and $(x, r+M) \in \text{epi}f_2$. □

The following two lemmas are contained in Volle [15, p. 848]. They are needed for obtaining the subdifferential formula for $f_+ = \max\{f, 0\}$ in Lemma 2.6.

LEMMA 2.4 (see [15, p. 848]). *Let C_1 and C_2 be closed convex sets in X and $C_1 \cap C_2 \neq \emptyset$. If $\text{cone}(C_1 - C_2)$ is a closed subspace, then for all $x \in C_1 \cap C_2$, $N_{C_1 \cap C_2}(x) = N_{C_1}(x) + N_{C_2}(x)$.*

LEMMA 2.5 (see [15, p. 848]). *Let f_1 and f_2 be proper convex functions on X, $g = \max\{f_1, f_2\}$, and $a \in \text{dom}f_1 \cap \text{dom}f_2$. If $f_1(a) = f_2(a)$, $\partial g(a) \neq \emptyset$, and $N_{\text{epi}g}(a, g(a)) = N_{\text{epi}f_1}(a, f_1(a)) + N_{\text{epi}f_2}(a, f_2(a))$, then $\partial g(a) = \text{conv}(\partial f_1(a) \cup \partial f_2(a)) + N_{\text{dom}f_1}(a) + N_{\text{dom}f_2}(a)$.*

In the next lemma, we express $\partial f_+(x)$ in terms of $\partial f(x)$ and $\partial^\infty f(x)$ when $f(x) = 0$. Note that because f is not necessarily lower semicontinuous, $\text{epi}f$ may not be closed.

LEMMA 2.6. *If $f(x) = 0$, then $\partial f_+(x) = [0, 1]\partial f(x) + \partial^\infty f(x)$.*

Proof. Let $f_1 = f$, $f_2 = 0$, and $g(z) = \max\{f(z), 0\} = f_+(z)$. Then $\text{dom}f_2 = X$, $\text{cone}(\text{dom}f_1 - \text{dom}f_2) = X$, and $\text{epi}f_2 = X \times R_+$. It follows from Lemma 2.3(ii) that

$$\begin{aligned} X \times R &= \text{cone}(\text{dom}f_1 - \text{dom}f_2) \times R \\ &= \text{cone}(\text{epi}f_1 - \text{epi}f_2) \\ &\subseteq \text{cone}(\overline{\text{epi}f_1} - \text{epi}f_2) \\ &\subseteq X \times R. \end{aligned}$$

Therefore, $\text{cone}(\overline{\text{epi}f_1} - \text{epi}f_2) = X \times R$ is a closed subspace of $X \times R$. Because $f(x) = 0$, $(x, 0) \in \overline{\text{epi}f_1} \cap \text{epi}f_2$. From Lemma 2.4,

$$N_{\overline{\text{epi}f_1} \cap \text{epi}f_2}(x, 0) = N_{\overline{\text{epi}f_1}}(x, 0) + N_{\text{epi}f_2}(x, 0). \tag{2.2}$$

From Lemma 2.3(i), $\overline{\text{epi}f_1} \cap \text{epi}f_2 = \overline{\text{epi}f_1 \cap \text{epi}f_2}$. Because $(x, 0) \in \text{epi}f_1 \cap \text{epi}f_2$, we have

$$N_{\overline{\text{epi}f_1 \cap \text{epi}f_2}}(x, 0) = N_{\text{epi}f_1 \cap \text{epi}f_2}(x, 0) \quad \text{and} \quad N_{\overline{\text{epi}f_1}}(x, 0) = N_{\text{epi}f_1}(x, 0).$$

Consequently, (2.2) is reduced to

$$N_{\text{epi}f_1 \cap \text{epi}f_2}(x, 0) = N_{\text{epi}f_1}(x, 0) + N_{\text{epi}f_2}(x, 0). \tag{2.3}$$

Because $f_1(x) = f_2(x) = 0$, $0 \in \partial g(x) \neq \emptyset$, and (2.3) holds, from Lemma 2.5 one has that

$$\begin{aligned} \partial g(x) &= \text{conv}(\partial f_1(x) \cup \partial f_2(x)) + N_{\text{dom}f_1}(x) + N_{\text{dom}f_2}(x) \\ &= \text{conv}(\partial f_1(x) \cup \{0\}) + N_{\text{dom}f_1}(x) + N_X(x) \\ &= [0, 1]\partial f(x) + \partial^\infty f(x). \end{aligned}$$ □

3. Weak BCQ and direction-length characterization of error bounds. In this section, we characterize local and global error bounds in terms of normal cones, subdifferentials, and the end set of subdifferentials.

It is well known that for a continuous convex function f, (1.1) is said to satisfy the basic constraint qualification (BCQ) at $x \in bd(S)$ if $N_S(x) = [0, \infty)\partial f(x)$ (see, e.g., [5, 8, 9]). As defined in [19], (1.1) is said to satisfy the extended BCQ at $x \in bd(S)$ if $N_S(x) = [0, \infty)\partial f(x) + \partial^\infty f(x)$, and (1.1) is said to satisfy the strong BCQ at

$x \in \mathrm{bd}(S)$ if there exists $\tau \in (0,\infty)$ such that $N_S(x) \cap B^* \subseteq [0,\tau]\partial f(x) + \partial^\infty f(x)$, where B^* denotes the closed unit ball of X^*.

DEFINITION 3.1. *Inequality* (1.1) *is said to satisfy the weak BCQ at* $x \in S$ *if*

$$N_S(x) \subseteq [0,\infty)\partial f(x) + \partial^\infty f(x).$$

Note that the nontrivial case is when the point x is on the boundary. For any $x \in \mathrm{int}S$, $N_S(x) = \{0\}$ and the weak BCQ always holds. If $x \in \mathrm{bd}(S)$ and $f(x) = 0$, then $\partial f(x) \subseteq N_S(x)$. Under this condition, the weak BCQ is equivalent to the extended BCQ.

Example 3.1. Consider the following proper convex function on R^1:

$$f(x) = \begin{cases} -x - 1 & \text{if } |x| \le 1; \\ \infty & \text{otherwise.} \end{cases} \tag{3.1}$$

It is not difficult to verify that $S = [-1,1]$, $f(1) < 0$,

$$N_S(1) = \{x^* : x^*(y-1) \le 0 \ \forall y \in [-1,1]\} = [0,\infty) = \partial^\infty f(1),$$

$$\partial f(1) = \{x^* : -y - 1 \ge -2 + x^*(y-1) \ \forall y \in [-1,1]\} = [-1,\infty),$$

$$E[\partial f(1)] = \{-1\}, \quad \text{and} \quad d_{E[\partial f(1)]\cap N_S(1)}(0) = \infty.$$

In this case, the extended BCQ does not hold at $x = 1$. It is also interesting to observe that $\partial f(1) \not\subseteq N_S(1)$. However, we still have $N_S(1) \subset [0,\infty)\partial f(1) + \partial^\infty f(1)$ and $d_{E[\partial f(1)]\cap N_S(1)}(0) = \infty \ge 1/\tau$ for all $\tau > 0$. It is obvious (also confirmed by Theorem 3.1 and [1, 4, 16, 17, 18]) that $f(x) \le 0$ has a local error bound τ at $x = 1$ for any $\tau > 0$. This is a case not covered by [6, 19].

Example 3.1 illustrates that the weak BCQ is weaker than the extended BCQ. It requires only that all normal directions are controlled by the cone $[0,\infty)\partial f(x) + \partial^\infty f(x)$. We are going to use the weak BCQ plus the length control of $E[\partial f(x)] \cap N_S(x)$ to characterize local error bounds. The following fact is useful for computing $E[\partial f(x)] \cap N_S(x)$.

FACT 3.1. $E[\partial f(x)] \cap N_S(x) = \{x^* \in \overline{[0,1]\partial f(x)} \cap N_S(x) : \ tx^* \notin \overline{[0,1]\partial f(x)} \cap N_S(x)$ *for all* $t > 1\}$. *Indeed, if* $x^* \in E[\partial f(x)] \cap N_S(x)$, *then* $x^* \in \overline{[0,1]\partial f(x)} \cap N_S(x)$. *As* $x^* \in E[\partial f(x)]$, $tx^* \notin \overline{[0,1]\partial f(x)}$ *for all* $t > 1$, *which implies* $tx^* \notin [0,1]\partial f(x) \cap N_S(x)$ *for all* $t > 1$. *On the other hand, given* $x^* \in \overline{[0,1]\partial f(x)} \cap N_S(x)$ *and* $tx^* \notin \overline{[0,1]\partial f(x)} \cap N_S(x)$ *for all* $t > 1$, *since* $tx^* \in N_S(x)$, *one has* $tx^* \notin \overline{[0,1]\partial f(x)}$ *for all* $t > 1$. *Thus,* $x^* \in E[\partial f(x)] \cap N_S(x)$.

THEOREM 3.1. *Let* $a \in \mathrm{bd}(S)$ *and* $\tau \in (0,\infty)$. *Then* (1.1) *has a local error bound* τ *at* a *if and only if there exists* $r \in (0,\infty)$ *such that for all* $x \in B(a,r) \cap \mathrm{bd}(S)$, (A) *the weak BCQ holds at* x, *and* (B) $d_{E[\partial f(x)]\cap N_S(x)}(0) \ge \tau^{-1}$.

Proof. " $\Rightarrow$ " If (1.1) has a local error bound τ at a, there exists $r > 0$ such that $d_S(x) \le \tau f_+(x)$ for all $x \in B(a,r)$. For any $x \in B(a,r) \cap \mathrm{bd}(S)$, we discuss two cases: $f(x) = 0$ and $f(x) < 0$. In the first case of $f(x) = 0$, let $x^* \in \partial d_S(x)$. Because $d_S(\cdot)$ is

a continuous convex function, by Fact 2.2 and $x \in B(a,r)$, one has that for all $h \in X$,

$$\begin{aligned}\langle x^*, h\rangle &\le \sup\{\langle y^*, h\rangle : y^* \in \partial d_S(x)\}\\ &= \lim_{t\to 0^+} \frac{d_S(x+th) - d_S(x)}{t}\\ &\le \lim_{t\to 0^+} \frac{\tau f_+(x+th) - \tau f_+(x)}{t}\\ &= \tau f'_+(x;h).\end{aligned}$$

It follows from Fact 2.2 that $x^* \in \tau\partial f_+(x)$, and thus $\partial d_S(x) \subseteq \tau\partial f_+(x)$. By Lemma 2.6,

$$N_S(x) \cap B^* = \partial d_S(x) \subseteq \tau([0,1]\partial f(x) + \partial^\infty f(x)) = [0,\tau]\partial f(x) + \partial^\infty f(x), \tag{3.2}$$

which implies (A). Now we prove (B). Without loss of generality we may assume that $E[\partial f(x)] \cap N_S(x) \neq \emptyset$, i.e., $d_{E[\partial f(x)]\cap N_S(x)}(0) < \infty$. Given $x^* \in E[\partial f(x)] \cap N_S(x)$, by the definition of the end set, we have $x^* \neq 0$. By (3.2), there exist $\mu \in [0,\tau]$, $y^* \in \partial f(x)$, and $z^* \in \partial^\infty f(x)$ satisfying $\frac{x^*}{\|x^*\|} = \mu y^* + z^*$. If $\mu = 0$, then $x^* = \|x^*\|z^* \in \partial^\infty f(x)$. By Fact 2.4 and the fact that $\partial^\infty f(x)$ is a cone, $[0,\infty)x^* \subseteq \partial^\infty f(x) \subseteq \overline{[0,1]\partial f(x)}$, which contradicts $x^* \in E[\partial f(x)]$. Hence, $\mu > 0$ and

$$\frac{x^*}{\mu\|x^*\|} = y^* + \mu^{-1}z^* \in \partial f(x) + \mu^{-1}\partial^\infty f(x) = \partial f(x) \subseteq \overline{[0,1]\partial f(x)}.$$

Since $x^* \in E[\partial f(x)]$, by the definition of the end set, we must have $\frac{1}{\mu\|x^*\|} \le 1$, which implies that $\|x^*\| \ge \mu^{-1} \ge \tau^{-1}$. Therefore, $d_{E[\partial f(x)]\cap N_S(x)}(0) \ge \tau^{-1}$, which is (B).

In the second case of $f(x) < 0$, by Lemma 2.2, $N_S(x) = \partial^\infty f(x) \subseteq [0,\infty)\partial f(x) + \partial^\infty f(x)$, and thus (A) holds. We show that $d_{E[\partial f(x)]\cap N_S(x)}(0) = \infty$, i.e., $E[\partial f(x)] \cap N_S(x) = \emptyset$. Indeed, if $\partial f(x) = \emptyset$, then $E[\partial f(x)] = \emptyset$ and thus $E[\partial f(x)] \cap N_S(x) = \emptyset$. If $\partial f(x) \neq \emptyset$ and $f(x) < 0$, from Lemma 2.2 and Fact 2.4, $N_S(x) = \partial^\infty f(x) \subseteq \overline{[0,1]\partial f(x)}$, which implies $\overline{[0,1]\partial f(x)} \cap N_S(x) = N_S(x)$. By Fact 3.1, one has that $E[\partial f(x)] \cap N_S(x) = \{x^* \in N_S(x) : \ tx^* \notin N_S(x) \text{ for all } t > 1\} = \emptyset$, which implies $d_{E[\partial f(x)]\cap N_S(x)}(0) = \infty$.

" $\Leftarrow$ " We show that $d_S(x) \le \tau f_+(x)$ for all $x \in B(a, r/2)$. If $x \in S$ or $f_+(x) = \infty$, then $d_S(x) \le \tau f_+(x)$ holds. Now suppose that $x \notin S$ and $f_+(x) < \infty$. Given $x \in B(a,r/2)$, $d_S(x) \le \|x - a\| < r/2$. Choose $t \in (0,1)$ such that $d_S(x) < t(r/2)$. By [10, Lemma 1.1 and Proposition 1.3], for this t there exist $x_t \in \mathrm{bd}(S)$ and $x_t^* \in N_S(x_t) \cap B^* = \partial d_S(x_t)$ satisfying $t\|x - x_t\| \le \langle x_t^*, x - x_t\rangle$. Because $x_t^* \in \partial d_S(x_t)$ and $x_t \in \mathrm{bd}(S)$, one has

$$t\|x - x_t\| \le \langle x_t^*, x - x_t\rangle \le d_S(x) - d_S(x_t) = d_S(x) < t(r/2). \tag{3.3}$$

Hence, $\|x_t - a\| \le \|x_t - x\| + \|x - a\| < r/2 + r/2 = r$, and $x_t \in B(a,r) \cap \mathrm{bd}(S)$. We claim that

$$\partial d_S(x) \subseteq \tau\partial f_+(x) \quad \forall x \in B(a,r) \cap \mathrm{bd}(S). \tag{3.4}$$

Indeed, let $x \in B(a,r) \cap \mathrm{bd}(S)$ and $x^* \in \partial d_S(x) = N_S(x) \cap B^*$. Since $f_+(x) = 0$ and x minimizes f_+, $0 \in \tau\partial f_+(x)$. Thus we only need to show $x^* \in \tau\partial f_+(x)$ for $x^* \neq 0$. If $f(x) < 0$, by Lemma 2.2, $x^* \in N_S(x) = \partial f_+(x) = \tau\partial f_+(x)$. If $f(x) = 0$ and $\partial f(x) = \emptyset$, by (A) and Lemma 2.6, $x^* \in N_S(x) \subseteq \partial^\infty f(x) = [0,\tau]\partial f(x) + \partial^\infty f(x) =$

$\tau\partial f_+(x)$. If $f(x) = 0$ and $\partial f(x) \neq \emptyset$, by (A), there exist $\lambda \geq 0$, $u^* \in \partial f(x)$, and $v^* \in \partial^\infty f(x)$ satisfying $x^* = \lambda u^* + v^*$. If $\lambda = 0$, then by Lemma 2.6, $x^* = v^* \in \partial^\infty f(x) \subseteq [0,\tau]\partial f(x) + \partial^\infty f(x) = \tau\partial f_+(x)$. Otherwise,

$$\lambda^{-1}x^* = u^* + \lambda^{-1}v^* \in \partial f(x) + \lambda^{-1}\partial^\infty f(x) = \partial f(x) \subseteq \overline{[0,1]\partial f(x)}.$$

Let $M = \sup\{\alpha \geq 0 : \alpha x^* \in \overline{[0,1]\partial f(x)}\} \geq \lambda^{-1}$. If $M < \infty$, there exists an increasing sequence $\{t_n\}$ such that $\lim_{n\to\infty} t_n = M$, $t_n x^* \in \overline{[0,1]\partial f(x)}$, and $\lim_{n\to\infty} t_n x^* = Mx^* \in \overline{[0,1]\partial f(x)}$. By the definition of M, for all $t > 1$, $tMx^* \notin \overline{[0,1]\partial f(x)}$. Therefore, $Mx^* \in E[\partial f(x)]$. By (B),

$$M \geq \|Mx^*\| \geq d_{E[\partial f(x)]\cap N_S(x)}(0) \geq \frac{1}{\tau}.$$

Consequently, by Fact 2.4 and Lemma 2.6,

$$\begin{aligned} x^* = M^{-1}Mx^* &\in M^{-1}\overline{[0,1]\partial f(x)} \\ &= M^{-1}([0,1]\partial f(x) + \partial^\infty f(x)) \\ &\subseteq [0,\tau]\partial f(x) + \partial^\infty f(x) \\ &= \tau\partial f_+(x). \end{aligned}$$

If $M = \infty$, then by the convexity of $\overline{[0,1]\partial f(x)}$ and Fact 2.4,

$$[0,\infty)x^* \subseteq \overline{[0,1]\partial f(x)} = [0,1]\partial f(x) + \partial^\infty f(x).$$

By Lemma 2.6,

$$x^* = \tau\tau^{-1}x^* \in [0,\tau][0,\infty)x^* \subseteq [0,\tau]\partial f(x) + \partial^\infty f(x) = \tau\partial f_+(x).$$

Thus (3.4) is proved, and we have $x_t^* \in \partial d_S(x_t) \subseteq \tau\partial f_+(x_t)$. It follows from (3.3) and $x_t^* \in \tau\partial f_+(x_t)$ that

$$td_S(x) \leq t\|x - x_t\| \leq \langle x_t^*, x - x_t\rangle \leq \tau f_+(x) - \tau f_+(x_t) = \tau f_+(x). \tag{3.5}$$

Since (3.5) holds for all t sufficiently close to 1, one has $d_S(x) \leq \tau f_+(x)$. □

Remark 3.1. (i) For $x \in \mathrm{int}(S)$, $N_S(x) = \{0\}$ and $E[\partial f(x)] \cap N_S(x) = \emptyset$. Thus, (A), (B), and the strong BCQ hold. For $x \in \mathrm{bd}(S)$ and $f(x) < 0$, as indicated in the proof of Theorem 3.1, (A), (B), and the strong BCQ hold by Lemma 2.2. For $x \in \mathrm{bd}(S)$ and $f(x) = 0$, the strong BCQ (3.2) implies (A) and (B), and (A) and (B) imply (3.4), which is the strong BCQ. Therefore, the strong BCQ is pointwise equivalent to (A) and (B), and it characterizes local error bounds without the boundary restriction $\mathrm{bd}(S) \subseteq f^{-1}(0)$ and lower semicontinuity. This extends [6, Corollary 4.1] and [19, Theorem 2.2]. For a discontinuous function, the removal of the boundary restriction $\mathrm{bd}(S) \subseteq f^{-1}(0)$ is nontrivial.

(ii) When $x \in \mathrm{bd}(S)$ and $f(x) = 0$, the weak BCQ is equivalent to the extended BCQ and $E[\partial f(x)] \cap N_S(x) = E[\partial f(x)]$. Under this condition, (A) and (B) recover the decomposition of the strong BCQ in [6, Theorem 4.1].

(iii) When f is lower semicontinuous, the sufficient part of Theorem 3.1 can be derived from [19, Theorem 2.2], [6, Theorem 4.1], [13].

Theorem 3.1 explains that the nonexistence of a local error bound is possibly caused by either of the following: the cone $[0,\infty)\partial f(x) + \partial^\infty f(x)$ fails to represent all normal directions or the set $E[\partial f(x)] \cap N_S(x)$ is not separated from the origin. When

(A) and (B) hold uniformly on the entire boundary, we obtain a direction-length characterization for global error bounds.

THEOREM 3.2. *Let $\tau \in (0,\infty)$. Then $d_S(x) \le \tau f_+(x)$ for all $x \in X$ if and only if the weak BCQ and $d_{E[\partial f(x)]\cap N_S(x)}(0) \ge \tau^{-1}$ hold on* $\mathrm{bd}(S)$.

Proof. If $d_S(x) \le \tau f_+(x)$ for all $x \in X$, then (1.1) has a local error bound τ at every point of $\mathrm{bd}(S)$. By Theorem 3.1, the weak BCQ and $d_{E[\partial f(x)]\cap N_S(x)}(0) \ge \frac{1}{\tau}$ hold for all $x \in \mathrm{bd}(S)$. Conversely, for any $z \in X$, there exist $a \in \mathrm{bd}(S)$ and $r \in (0,\infty)$ such that $z \in B(a, \frac{r}{2})$. Since the weak BCQ and $d_{E[\partial f(x)]\cap N_S(x)}(0) \ge \frac{1}{\tau}$ hold for all $x \in B(a,r)\cap \mathrm{bd}(S)$, by the proof of Theorem 3.1, $d_S(x) \le \tau f_+(x)$ for all $x \in B(a, \frac{r}{2})$, and thus $d_S(z) \le \tau f_+(z)$. □

Note that the Slater condition implies the existence of a local error bound for any $x \in \mathrm{bd}(S)$ [13]. We have the following corollary.

COROLLARY 3.1. *If (1.1) satisfies the Slater condition, then for any $a \in \mathrm{bd}(S)$ there exist $r, \tau \in (0,\infty)$ such that for all $x \in B(a,r)\cap \mathrm{bd}(S)$,*

$$N_S(x) \subseteq [0,\infty)\partial f(x) + \partial^\infty f(x) \quad \textit{and} \quad d_{E[\partial f(x)]\cap N_S(x)}(0) \ge \tau^{-1}.$$

In the following examples, we use Theorems 3.1 and 3.2 to determine the existence or nonexistence of error bounds and to compute the smallest error bound if it exists.

Example 3.2. Let $\alpha > 1$, let x_i denote the ith component of $x \in R^n$, and let $g(x) = \frac{1}{\alpha}x_1$, $h(x) = \|x\| = (x_1^2 + \cdots + x_n^2)^{\frac{1}{2}}$, and $f(x) = g(x) + h(x)$. It is easy to verify that $S = \{0\}$ and $N_S(0) = R^n$. Since $ri(\mathrm{dom}(g)) \cap ri(\mathrm{dom}(h)) \ne \emptyset$, we have [14, Theorem 23.8]

$$\begin{aligned}\partial f(0) &= \partial g(0) + \partial h(0)\\ &= \{(\alpha^{-1}, 0, \ldots, 0)\} + \overline{B(0,1)}\\ &= \{x \in R^n : (x_1 - \alpha^{-1})^2 + x_2^2 + \cdots + x_n^2 \le 1\}.\end{aligned}$$

Thus, $E[\partial f(0)] = \{x \in R^n : (x_1 - \alpha^{-1})^2 + x_2^2 + \cdots + x_n^2 = 1\} = E[\partial f(0)] \cap N_S(0)$. In this case, the weak BCQ holds at 0 and $d_{E[\partial f(0)]\cap N_S(0)}(0) = 1 - \alpha^{-1}$. By Theorem 3.2 (or by [1, 4, 17, 18]), $\tau = \frac{\alpha}{\alpha - 1}$ is the smallest global error bound.

Example 3.3. Let $g(x) = x_1$, $h(x) = \|x\|$, and $f(x) = g(x) + h(x)$. One can verify that $S = \{(t, 0, \ldots, 0) \in R^n : t \le 0\}$, $N_S(0) = \{(x_1, x_2, \ldots, x_n) \in R^n : x_1 \ge 0\}$, and $\partial f(0) = \partial g(0) + \partial h(0) = \{x \in R^n : (x_1 - 1)^2 + x_2^2 + \cdots + x_n^2 \le 1\}$. In this case, $[0,\infty)\partial f(0) + \partial^\infty f(0)$ fails to represent all normal directions. In addition, $E[\partial f(0)] = \{x \ne 0 : (x_1 - 1)^2 + x_2^2 + \cdots + x_n^2 = 1\}$, and thus $d_{E[\partial f(0)]\cap N_S(0)}(0) = 0$. By Theorem 3.1, $f(x) \le 0$ does not have a local error bound at 0. Note that this example (for $n = 2$) was first used in [7] to illustrate the nonexistence of a global error bound. The nonexistence a local error bound at 0 can be derived from [2].

Example 3.4. Let $D = \{(x,y) : 0 < x \le 1, 0 \le y \le 1\} \cup \{(0,y) : 0 \le y \le \frac{1}{2}\}$, and define $f(x,y) = I_D(x,y) + x$, where $I_D(x,y)$ is the indicator function of D. Note that D is convex but not closed, and f is proper convex but not lower semicontinuous. It is not difficult to verify that

$$S = \{(x,y) : f(x,y) \le 0\} = \left\{(0,y) : 0 \le y \le \frac{1}{2}\right\},$$

$$N_S\left(0, \frac{1}{2}\right) = \{(x,y) : y \ge 0\},$$

$$\partial^\infty f\left(0,\frac{1}{2}\right) = N_{\mathrm{dom} f}\left(0,\frac{1}{2}\right) = N_D\left(0,\frac{1}{2}\right) = \{(x,0) : x \le 0\},$$

$$\partial f\left(0,\frac{1}{2}\right) = \partial I_D\left(0,\frac{1}{2}\right) + \{(1,0)\} = N_D\left(0,\frac{1}{2}\right) + \{(1,0)\} = \{(x,0) : x \le 1\}.$$

Hence, $[0,\infty)\partial f(0,\frac{1}{2}) + \partial^\infty f(0,\frac{1}{2}) = \{(x,0) : x \in R\}$, and the weak BCQ does not hold at $(0,\frac{1}{2})$. By Theorem 3.1, $f(x,y) \le 0$ does not have a local error bound at $(0,\frac{1}{2})$.

4. Segment extension property and global error bounds. In section 3 we have seen that (1.1) has a local error bound at a if (A) and (B) uniformly hold in a boundary neighborhood of a. When (A) and (B) uniformly hold in the entire boundary of S, we immediately obtain a characterization of global error bounds. In this section, we study global error bounds from a new aspect. We show that the verification of the direction-length decomposition can be sufficiently carried out on a small subset instead of the entire boundary. For this purpose, we introduce a useful new concept, the segment extension property.

DEFINITION 4.1. *Let $C \ne \emptyset$ be a convex set in X and $P \subseteq C$. Then P is said to have the segment extension property with respect to C (property SE) if for any $x \in C$ there exist $p \in P$ and $t > 1$ such that $p + t(x-p) \in C$.*

Note that the definition of property SE is equivalent to the following: For any $x \in C$, there exist $p \in P$, $c \in C$, and $\lambda \in (0,1)$ such that $x = (1-\lambda)p + \lambda c$.

Example 4.1. Given a convex set $C \ne \emptyset$, if there exists $P \subseteq C$ satisfying property R, i.e., $C = P + C^\infty$, then P has property SE. In particular, C has property SE with respect to itself. This was first introduced in [11, 20].

Example 4.2. If P has property SE with respect to C, then P contains all extreme points of C (if there are any). Indeed, if $e \in \mathrm{ext}(C)$ but $e \notin P$, then there exist $p \in P$ and $t > 1$ such that $x = p + t(e-p) \in C$. Hence, $e = (1-\frac{1}{t})p + \frac{1}{t}x$, which contradicts the definition of extreme point.

Example 4.3. If C is closed and $\emptyset \ne C \ne X$, then $\mathrm{bd}(C) \ne \emptyset$ has property SE.

In many cases a set having property SE could be much smaller than the entire boundary, as illustrated by the following examples.

Example 4.4. If C is a cone, then $P = \{0\}$ has property SE.

Example 4.5. Let $C = \{(x,y) : -\infty < x < \infty, -1 \le y \le 1\}$. Then $P = \{(0,1),(0,-1)\}$ has property SE.

Note that the convex set C in Example 4.5 has no extreme point. In general, let $L = C^\infty \cap (-C^\infty)$; then the set $P = \mathrm{ext}(C \cap L^\perp)$ has property SE, as proved in Lemma 4.1.

LEMMA 4.1. *Let $\emptyset \ne C \subset R^n$ be a closed convex set and $L = C^\infty \cap (-C^\infty)$. Then $\mathrm{ext}(C \cap L^\perp) \ne \emptyset$ has property SE. In particular, if $\mathrm{ext}(C) \ne \emptyset$, then $\mathrm{ext}(C)$ has property SE.*

Proof. First assume that C contains no lines. Then $L = \{0\}$, $L^\perp = R^n$, $C = C \cap L^\perp$, $\mathrm{ext}(C) \ne \emptyset$, and $C = \mathrm{conv}(\mathrm{ext}(C)) + C^\infty$ [14, pp. 166–167]. Given $x \in C$, there exist $e_i \in \mathrm{ext}(C)$ and $\lambda_i > 0$, $i = 1, \ldots, k$, $\sum_{i=1}^k \lambda_i = 1$, and $h \in C^\infty$ satisfying $x = \sum_{i=1}^k \lambda_i e_i + h$. If $k = 1$, then $e_1 + 2h = e_1 + 2(x - e_1) \in C$. If $k > 1$, let $\mu = \sum_{i=2}^k \lambda_i$ and $y = \sum_{i=2}^k \lambda_i e_i + h$. Then $\lambda_1 = 1 - \mu$, $\frac{1}{\mu} > 1$, $x = (1-\mu)e_1 + y$, and

$\mu e_1 + x - e_1 = y$. It follows that

$$e_1 + \frac{1}{\mu}(x - e_1) = \frac{1}{\mu}y = \sum_{i=2}^{k} \frac{\lambda_i}{\mu} e_i + \frac{1}{\mu}h \in C.$$

Therefore, $\text{ext}(C \cap L^{\perp}) = \text{ext}(C)$ has property SE.

In the case when C contains lines, the set $M = C \cap L^{\perp}$ contains no lines, $\text{ext}(M) \neq \emptyset$, and $C = M + L = \text{conv}(\text{ext}(M)) + M^{\infty} + L$ [14, p. 65]. The rest of the proof is similar to that of the first case and is omitted. □

In order to utilize property SE to reduce the verification of the weak BCQ and the length control of the subdifferential, we first study the relations between normal cones/subdifferentials at a point p and a point that can be extended from p.

LEMMA 4.2. *If C is a convex set in X, $a, b \in C$, $\lambda \in (0,1)$, and $x = (1-\lambda)a + \lambda b$, then $N_C(x) \subseteq N_C(a)$.*

Proof. Let $c = \lambda(b - a)$; then $x = a + c$ and $b = a + \frac{1}{\lambda}c$. For any $x^* \in N_C(x)$, $\langle x^*, a - x\rangle = \langle x^*, -c\rangle \leq 0$, which implies $\langle x^*, c\rangle \geq 0$. On the other hand, $\langle x^*, b - x\rangle = \langle x^*, (\frac{1}{\lambda} - 1)c\rangle \leq 0$, which implies $\langle x^*, c\rangle \leq 0$. Therefore, $\langle x^*, c\rangle = 0$. It follows that for all $u \in C$,

$$\langle x^*, u - a\rangle = \langle x^*, u - x + c\rangle = \langle x^*, u - x\rangle + \langle x^*, c\rangle = \langle x^*, u - x\rangle \leq 0.$$

Hence, $x^* \in N_C(a)$. □

LEMMA 4.3. *If $a, b \in S$, $\lambda \in (0,1)$, $x = (1-\lambda)a + \lambda b$, and $f(x) = f(a) = 0$, then $\partial f(x) = \partial f(a) \cap N_S(x)$.*

Proof. For any $x^* \in \partial f(x)$, $\langle x^*, u - x\rangle \leq f(u) - f(x)$ for all $u \in X$. By $f(x) = 0$ and Lemma 4.2, $\partial f(x) \subseteq N_S(x) \subseteq N_S(a)$, which implies $\langle x^*, x - a\rangle \leq 0$. Thus,

$$\langle x^*, u - a\rangle = \langle x^*, u - x\rangle + \langle x^*, x - a\rangle \leq \langle x^*, u - x\rangle \leq f(u) - f(x) = f(u) - f(a).$$

Therefore, $x^* \in \partial f(a)$ and thus $\partial f(x) \subseteq \partial f(a) \cap N_S(x)$. On the other hand, for $x^* \in \partial f(a) \cap N_S(x) \subseteq N_S(x)$, by the proof of Lemma 4.2, $\langle x^*, c\rangle = 0$, where $c = \lambda(b - a)$. Thus,

$$\begin{aligned} \langle x^*, u - x\rangle &= \langle x^*, u - x\rangle + \langle x^*, c\rangle \\ &= \langle x^*, u - x + c\rangle \\ &= \langle x^*, u - a\rangle \\ &\leq f(u) - f(a) \\ &= f(u) - f(x). \end{aligned}$$

Hence, $x^* \in \partial f(x)$ and $\partial f(a) \cap N_S(x) \subseteq \partial f(x)$. □

By using the above lemmas, we are ready to show that the weak BCQ and the length control of the subdifferential can be segment extended on S.

THEOREM 4.1. *Let $a, b \in S$, $\lambda \in (0,1)$, and $x = (1-\lambda)a + \lambda b$.*

(i) *If the weak BCQ holds at a, then it holds at x.*

(ii) *$E[\partial f(x)] \cap N_S(x) \subseteq E[\partial f(a)] \cap N_S(a)$, which implies that $d_{E[\partial f(x)] \cap N_S(x)}(0) \geq d_{E[\partial f(a)] \cap N_S(a)}(0)$.*

Proof. (i) If $x \notin \text{bd}(S)$, then $N_S(x) = \{0\} \subseteq [0,\infty)\partial f(x) + \partial^{\infty} f(x)$. If $x \in \text{bd}(S)$ and $f(x) < 0$, then by Lemma 2.2, $N_S(x) = \partial^{\infty} f(x) \subseteq [0,\infty)\partial f(x) + \partial^{\infty} f(x)$. If $x \in \text{bd}(S)$ and $f(x) = 0$, then by the convexity of f, we must have $f(a) = 0$. Since $\partial^{\infty} f(a)$ is a cone, by Lemma 2.6 we have

$$N_S(a) \subseteq [0,\infty)\partial f(a) + \partial^{\infty} f(a) = [0,\infty)([0,1]\partial f(a) + \partial^{\infty} f(a)) = [0,\infty)\partial f_+(a).$$

It follows that

$$\begin{aligned} N_S(x) &= N_S(a) \cap N_S(x) \quad \text{(by Lemma 4.2)} \\ &\subseteq [0,\infty)\partial f_+(a) \cap N_S(x) \\ &= [0,\infty)(\partial f_+(a) \cap N_S(x)) \\ &= [0,\infty)\partial f_+(x) \quad \text{(by Lemma 4.3)} \\ &= [0,\infty)\partial f(x) + \partial^\infty f(x) \quad \text{(by Lemma 2.6)}. \end{aligned}$$

It remains to prove (ii). Without loss of generality, we may assume that $\partial f(x) \neq \emptyset$. If $x \notin \mathrm{bd}(S)$, then $N_S(x) = \{0\}$. Because an end set cannot contain 0, $E[\partial f(x)] \cap N_S(x) = \emptyset$ and $d_{E[\partial f(x)]\cap N_S(x)}(0) = \infty$. If $x \in \mathrm{bd}(S)$ and $f(x) < 0$, then, as in the proof of Theorem 3.1, $E[\partial f(x)] \cap N_S(x) = \emptyset$ and $d_{E[\partial f(x)]\cap N_S(x)}(0) = \infty$. If $x \in \mathrm{bd}(S)$ and $f(x) = 0$, then by the convexity of f, we must have $f(a) = 0$. We claim that

$$E[\partial f(x)] \subseteq E[\partial f(a)]. \tag{4.1}$$

Indeed, since $f(x) = 0$ and $f(a) = 0$, by Fact 2.4 and Lemma 2.6, $\overline{[0,1]\partial f(x)} = \partial f_+(x)$ and $\overline{[0,1]\partial f(a)} = \partial f_+(a)$. For any $u^* \in E[\partial f(x)]$, $u^* \in \overline{[0,1]\partial f(x)} = \partial f_+(x)$ and $tu^* \notin \overline{[0,1]\partial f(x)} = \partial f_+(x)$ for all $t > 1$. By Lemma 4.3,

$$u^* \in \partial f_+(x) = \partial f_+(a) \cap N_S(x) \subseteq \partial f_+(a) = \overline{[0,1]\partial f(a)}$$

and

$$tu^* \notin \partial f_+(x) = \partial f_+(a) \cap N_S(x) \quad \forall t > 1.$$

Note that $tu^* \in N_S(x)$; we must have $tu^* \notin \partial f_+(a) = \overline{[0,1]\partial f(a)}$ for all $t > 1$. Consequently, $u^* \in E[\partial f(a)]$ by the definition of the end set, and (4.1) is proven. It follows from (4.1) and Lemma 4.2 that

$$E[\partial f(x)] \cap N_S(x) \subseteq E[\partial f(a)] \cap N_S(a),$$

which implies that

$$d_{E[\partial f(x)]\cap N_S(x)}(0) \geq d_{E[\partial f(a)]\cap N_S(a)}(0). \qquad \square$$

By Theorem 4.1 and the equivalent definition of a set with property SE, we immediately have the following theorem.

THEOREM 4.2. *Let $P \subseteq S$ have property SE with respect to S.*

(i) *The weak BCQ holds for all $x \in S$ if and only if it holds for all $x \in P$.*

(ii) *$d_{E[\partial f(x)]\cap N_S(x)}(0) \geq \frac{1}{\tau}$ holds for all $x \in S$ if and only if it holds for all $x \in P$.*

Combining Theorems 4.2 and 3.2, we immediately obtain a characterization of global error bounds on any set $P \subseteq \mathrm{bd}(S)$ having property SE with respect to S.

THEOREM 4.3. *Let $\tau \in (0, +\infty)$ and $P \subseteq \mathrm{bd}(S)$ have property SE with respect to S. Then $d_S(x) \leq \tau f_+(x)$ for all $x \in X$ if and only if the weak BCQ and $d_{E[\partial f(x)]\cap N_S(x)}(0) \geq \frac{1}{\tau}$ hold on P.*

Note that when $x \in \mathrm{bd}(S) \cap f^{-1}(0)$, $\partial f(x) \subseteq N_S(x)$. Therefore, $E[\partial f(x)] \subseteq N_S(x)$ and $d_{E[\partial f(x)]\cap N_S(x)}(0) = d_{E[\partial f(x)]}(0)$. When $x \in \mathrm{bd}(S)$ and $f(x) < 0$, $d_{E[\partial f(x)]\cap N_S(x)}(0) = \infty$. Therefore, we obtain the following formula for the smallest global error bound.

COROLLARY 4.1. *Let $P \subseteq \mathrm{bd}(S)$ have property SE and $\gamma = \inf\{d_{E[\partial f(x)]}(0) : x \in P \cap f^{-1}(0)\}$. If the weak BCQ holds on P and $\gamma \in (0, \infty)$, then $\tau = \frac{1}{\gamma}$ is the smallest*

global error bound. If the weak BCQ holds on P and $\gamma = \infty$, then every $\tau \in (0, \infty)$ is a global error bound.

Finally, for the important special case of $X = R^n$, by Lemma 4.1, we obtain a characterization of global error bounds by conditions on extreme points.

THEOREM 4.4. *Let $\tau \in (0, +\infty)$, $X = R^n$, $L = S^\infty \cap (-S^\infty)$, and $\gamma = \inf\{d_{E[\partial f(x)]}(0) : x \in \operatorname{ext}(S \cap L^\perp) \cap f^{-1}(0)\}$. Then $d_S(x) \le \tau f_+(x)$ for all $x \in X$ if and only if the weak BCQ and $d_{E[\partial f(x)] \cap N_S(x)}(0) \ge \frac{1}{\tau}$ hold on $\operatorname{ext}(S \cap L^\perp)$. If a global error bound exists and $\gamma \in (0, \infty)$, then $\tau = \frac{1}{\gamma}$ is the smallest global error bound.*

Acknowledgments. The author wishes to express her sincere appreciation to the referees and Prof. Jiri Outrata for their careful reading of this paper and valuable comments.

REFERENCES

[1] D. AZÉ AND J.-N. CORVELLEC, *On the sensitivity analysis of Hoffman constants for systems of linear inequalities*, SIAM J. Optim., 12 (2002), pp. 913–927.
[2] D. AZÉ AND J.-N. CORVELLEC, *Characterizations of error bounds for lower semicontinuous functions on metric spaces*, ESAIM Control Optim. Calc. Var., 10 (2004), pp. 409–425.
[3] P. L. COMBETTES, *Strong convergence of block-iterative outer approximation methods for convex optimization*, SIAM J. Control Optim., 38 (2000), pp. 538–565.
[4] O. CORNEJO, A. JOURANI, AND C. ZĂLINESCU, *Conditioning and upper-Lipschitz inverse subdifferentials in nonsmooth optimization problems*, J. Optim. Theory Appl., 95 (1997), pp. 127–148.
[5] J.-B. HIRIART-URRUTY AND C. LEMARÉCHAL, *Convex Analysis and Minimization Algorithms* I, Springer-Verlag, Berlin, Heidelberg, 1993.
[6] H. HU, *Characterizations of the strong basic constraint qualifications*, Math. Oper. Res., 30 (2005), pp. 956–965.
[7] A. LEWIS AND J. PANG, *Error bounds for convex inequality systems*, Nonconvex Optim. Appl. 27 (1997), pp. 75–100.
[8] W. LI, *Abadie's constraint qualification, metric regularity, and error bounds for differentiable convex inequalities*, SIAM J. Optim., 7 (1997), pp. 966–978.
[9] W. LI, C. NAHAK, AND I. SINGER, *Constraint qualifications for semi-infinite systems of convex inequalities*, SIAM J. Optim., 11 (2000), pp. 31–52.
[10] K. F. NG AND W. H. YANG, *Error bounds for abstract linear inequality systems*, SIAM J. Optim., 13 (2002), pp. 24–43.
[11] K. NG AND X. ZHENG, *Characterizations of error bounds for convex multifunctions on Banach spaces*, Math. Oper. Res., 29 (2004), pp. 45–63.
[12] J. PANG, *Error bounds in mathematical programming*, Math. Program., 79 (1997), pp. 299–332.
[13] S. M. ROBINSON, *An application of error bounds for convex programming in a linear space*, SIAM J. Control Optim., 13 (1975), pp. 271–273.
[14] R. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[15] M. VOLLE, *Sous-différentiel d'une enveloppe supérieure de fonctions convexes*, C. R. Acad. Sci. Paris Sér. I Math., 317 (1993), pp. 845–849.
[16] Z. WU AND J. J. YE, *First-order and second-order conditions for error bounds*, SIAM J. Optim., 14 (2003), pp. 621–645.
[17] C. ZĂLINESCU, *Weak sharp minima, well behaving functions, and global error bounds for convex inequalities in Banach spaces*, in Proceedings of 12th Baikal International School-Seminar on Optimization Methods and Their Applications, V. Bulatov and V. Baturin, eds., Irkutsk, 2001, pp. 272–284.
[18] C. ZĂLINESCU, *Convex Analysis in General Vector Spaces*, World Scientific, River Edge, NJ, 2002.
[19] X. Y. ZHENG AND K. F. NG, *Metric regularity and constraint qualifications for convex inequalities on Banach spaces*, SIAM J. Optim., 14 (2004), pp. 757–772.
[20] X. ZHENG, *Error bounds for set inclusions*, Sci. China Ser. A, 4 (2003), pp. 750–763.

SIAM J. OPTIM.
Vol. 18, No. 1, pp. 322–337

STABILITY AND SENSITIVITY OF OPTIMIZATION PROBLEMS WITH FIRST ORDER STOCHASTIC DOMINANCE CONSTRAINTS*

DARINKA DENTCHEVA†, RENÉ HENRION‡, AND ANDRZEJ RUSZCZYŃSKI§

Abstract. We analyze the stability and sensitivity of stochastic optimization problems with stochastic dominance constraints of first order. We consider general perturbations of the underlying probability measures in the space of regular measures equipped with a suitable discrepancy distance. We show that the graph of the feasible set mapping is closed under rather general assumptions. We obtain conditions for the continuity of the optimal value and upper-semicontinuity of the optimal solutions, as well as quantitative stability estimates of Lipschitz type. Furthermore, we analyze the sensitivity of the optimal value and obtain upper and lower bounds for the directional derivatives of the optimal value. The estimates are formulated in terms of the dual utility functions associated with the dominance constraints.

Key words. stochastic programming, stochastic ordering, semi-infinite optimization, chance constraints, Lipschitz stability, metric regularity, directional differentiability

AMS subject classifications. Primary, 90C15, 90C34, 90C48; Secondary, 46N10, 60E15, 91B06

DOI. 10.1137/060650118

1. Introduction.

The notion of stochastic ordering (or *stochastic dominance of first order*) was introduced in statistics in [14, 13] and further applied and developed in economics [17, 7, 6]. It is defined as follows. For a random variable X we consider its distribution function, $F(X;\eta) = P[X \le \eta]$, $\eta \in \mathbb{R}$. We say that a random variable X *dominates in the first order* a random variable Y if

$$F(X;\eta) \le F(Y;\eta) \quad \forall\, \eta \in \mathbb{R}. \tag{1.1}$$

We denote this relation $X \succeq_{(1)} Y$. For a modern perspective on stochastic orders, see [15, 25].

Let $g : \mathbb{R}^n \times \mathbb{R}^s \to \mathbb{R}$ be continuous with respect to both arguments, and let V be an s-dimensional random vector, defined on a certain probability space $(\Omega, \mathscr{F}, P)$. For every $z \in \mathbb{R}^n$

$$X_z(\omega) = g(z, V(\omega)), \quad \omega \in \Omega,$$

is a random variable. Given a benchmark random variable Y (defined on the same probability space), an optimization model with first order stochastic dominance constraint is formulated as follows:

$$\begin{aligned} &\min\ f(z) \\ &\text{s.t. } X_z \succeq_{(1)} Y, \\ &\quad z \in Z, \end{aligned} \tag{1.2}$$

*Received by the editors January 16, 2006; accepted for publication (in revised form) September 12, 2006; published electronically April 17, 2007. This work was supported by NSF awards DMS-0303545, DMS-0303728, DMI-0354500, DMI-0354678, DMS-0603728, DMS-0604060, and by the DFG Research Center MATHEON *Mathematics for Key Technologies* in Berlin.
http://www.siam.org/journals/siopt/18-1/65011.html

†Department of Mathematical Sciences, Stevens Institute of Technology, Castle Point on Hudson, Hoboken, NJ 07030 (ddentche@stevens.edu).

‡Weierstrass Institute, Mohrenstr. 39, 10117 Berlin, Germany (henrion@wias-berlin.de).

§Department of Management Science and Information Systems, Rutgers University, 94 Rockafeller Rd., Piscataway, NJ 08854 (rusz@business.rutgers.edu).

where $f : \mathbb{R}^n \to \mathbb{R}$ and $Z \subset \mathbb{R}^n$. Using definition (1.1), we can express the dominance constraint as a continuum of probabilistic constraints:

$$P\big[g(z,V) \geq \eta\big] \geq P\big[Y \geq \eta\big], \quad \eta \in \mathbb{R}.$$

In [5] optimality conditions for a relaxation of problem (1.2) were investigated, in which the dominance constraint was enforced on an interval $[a,b]$ rather than on the entire real line:

$$\begin{aligned} &\min\ f(z) \\ &\text{s.t. } P\big[g(z,V) \geq \eta\big] \geq P\big[Y \geq \eta\big], \quad \eta \in [a,b], \\ &\quad z \in Z. \end{aligned} \tag{1.3}$$

The restriction of the range of η to a compact interval is motivated by the need to satisfy a constraint qualification condition for the problem (see Definition 2.4). Both probability functions in problem (1.3) converge to 0 when $\eta \to \infty$ and to 1 when $\eta \to \infty$, which precludes Robinson-type conditions on the whole real line.

From now on, we shall assume that f is continuous and Z is a nonempty closed convex set. Our objective is to investigate the stability and sensitivity of the optimal value, the feasible set, and solution set, respectively, of problem (1.3) when the random variables V and Y are subject to perturbations.

For the purpose of our analysis it is convenient to formulate the dominance constraint with the use of "$\geq$" inequalities, as in (1.3). When the distributions are continuous, this formulation is equivalent to the formulation used in [5].

Problems with stochastic dominance constraints are new optimization models involving risk aversion (see [3, 4, 5]). As problems with a continuum of constraints on probability, they pose specific analytical and computational challenges. The probabilistic nature of the problem prevents the direct application of the theory of semi-infinite optimization. On the other hand, the specific structure of dominance constraints is significantly different from the structure of finitely many probabilistic constraints. Our stability analysis follows similar patterns to those in [8, 22, 23], where the focus was on probabilistic constraints. However, a straightforward application of those results (a recent overview of which can be found in [21]) is not possible due to the specific structure of problem (1.3). First, in (1.3) we deal with two separate probability terms due to the consideration of a benchmark variable. Second, and more importantly, problem (1.3) has a continuum of constraints which requires a more sophisticated analysis than the case of a finite family of constraints.

In section 2, we establish the closedness of the feasible set mapping, and we obtain stability results for the optimal value, for the feasible set, and for the solution set. In section 3, we analyze the sensitivity of the optimal value function, and we obtain bounds for its directional derivatives.

2. Stability. It is obvious from the formulation of the dominance constraint that only the distribution laws of V and Y matter there. Therefore, we introduce the measures μ_0 on $\mathbb{R}^s$ and ν_0 on $\mathbb{R}$ induced by V and Y. For all Borel sets $A \subset \mathbb{R}^s$ and $B \subset \mathbb{R}$,

$$\begin{aligned} \mu_0(A) &= P[V \in A], \\ \nu_0(B) &= P[Y \in B]. \end{aligned}$$

We denote the set of probability measures on $\mathbb{R}^m$ by $\mathscr{P}(\mathbb{R}^m)$.

Furthermore, we introduce the multifunction $H : \mathbb{R}^n \times \mathbb{R} \rightrightarrows \mathbb{R}^s$ defined by

$$H(z,\eta) := \{v \in \mathbb{R}^s : g(z,v) \geq \eta\}.$$

We consider the following parametric optimization problem:

$$\begin{aligned} &\min\ f(z) \\ &\text{s.t. } \mu(H(z,\eta)) - \nu([\eta,\infty)) \geq 0 \quad \forall \eta \in [a,b], \\ &\quad z \in Z, \end{aligned} \tag{2.1}$$

with parameters $\mu \in \mathscr{P}(\mathbb{R}^s)$ and $\nu \in \mathscr{P}(\mathbb{R})$. The original problem (1.3) is obtained when $(\mu,\nu) = (\mu_0,\nu_0)$. Our aim is to study the stability of solutions and of the optimal value to (2.1) under small perturbations of the underlying distributions μ_0 and ν_0.

For this purpose we equip the space $\mathscr{P}(\mathbb{R})$ with the Kolmogorov distance function:

$$\alpha_1(\nu_1,\nu_2) = \sup_{\eta\in\mathbb{R}} \left|\nu_1([\eta,\infty)) - \nu_2([\eta,\infty))\right|.$$

To introduce a distance function on $\mathscr{P}(\mathbb{R}^s)$, which is appropriate for our problem, we define the family of sets:

$$\mathscr{B} := \{H(z,\eta) : z \in Z, \eta \in [a,b]\} \cup \{v + \mathbb{R}^s_- : v \in \mathbb{R}^s\}.$$

The distance function on $\mathscr{P}(\mathbb{R}^s)$ is defined as the discrepancy

$$\alpha_{\mathrm{B}}\,(\mu_1,\mu_2) := \sup_{B\in\mathrm{B}} \left|\mu_1(B) - \mu_2(B)\right|.$$

On the product space $\mathscr{P}(\mathbb{R}^s) \times \mathscr{P}(\mathbb{R})$ we introduce the natural distance:

$$\alpha((\mu_1,\nu_1),(\mu_2,\nu_2)) := \max\{\alpha_{\mathrm{B}}\,(\mu_1,\mu_2), \alpha_1(\nu_1,\nu_2)\}. \tag{2.2}$$

Note that α is a metric, because the measures are compared, in particular, on all the cells of form $z + \mathbb{R}^s_-$ and $(-\infty,\eta)$, respectively.

We consider the constraint set mapping $\Phi : \mathscr{P}(\mathbb{R}^s) \times \mathscr{P}(\mathbb{R}) \rightrightarrows \mathbb{R}^n$, which assigns to every parameter (μ,ν) the feasible set of problem (2.1), i.e.,

$$\Phi(\mu,\nu) := \big\{z \in Z : \mu(H(z,\eta)) - \nu([\eta,\infty)) \geq 0 \quad \forall \eta \in [a,b]\,\big\}.$$

Given any open subset $U \subseteq \mathbb{R}^n$, we define the U-localized optimal value function, $\varphi_U : \mathscr{P}(\mathbb{R}^s) \times \mathscr{P}(\mathbb{R}) \to \overline{\mathbb{R}}$, of problem (2.1) as follows:

$$\varphi_U(\mu,\nu) := \inf\big\{f(z) : z \in \Phi(\mu,\nu) \cap \operatorname{cl} U\big\}.$$

The U-localized solution set mapping $\Psi_U : \mathscr{P}(\mathbb{R}^s) \times \mathscr{P}(\mathbb{R}) \rightrightarrows \mathbb{R}^n$ of problem (2.1) is defined by

$$\Psi_U(\mu,\nu) := \big\{z \in \Phi(\mu,\nu) \cap \operatorname{cl} U : f(z) = \varphi_U(\mu,\nu)\big\}.$$

When $U = \mathbb{R}^n$ we simply write $\varphi(\mu,\nu)$ and $\Psi(\mu,\nu)$.

The reason to consider localized mappings is that we allow general perturbations of the probability distributions. Then, without additional compactness conditions, no reasonable constraint qualification formulated at the solution points of the original problem (1.3) could guarantee stability of the global solution set mapping $\Psi := \Psi_{\mathbb{R}^n}$.

We recall a general stability result from [10, Proposition 1 and Theorem 1] in a version adapted to our setting. In the theorem below, the symbol $\mathbb{B}(z,r)$ denotes the ball about z of radius r.

THEOREM 2.1. *Let the following assumptions be satisfied in* (2.1):

1. *The original solution set $\Psi(\mu_0,\nu_0)$ is nonempty and bounded.*
2. *The graph of the constraint set mapping Φ is closed.*
3. *At every solution $z^0 \in \Psi(\mu_0,\nu_0)$ of the original problem, there exist $\varepsilon > 0$ and $L > 0$ such that for all $(\mu,\nu) \in \mathbb{B}((\mu_0,\nu_0);\varepsilon)$ the constraint set mapping satisfies the following two Lipschitz-like estimates:*

$$d(z,\Phi(\mu_0,\nu_0)) \le L\alpha((\mu,\nu),(\mu_0,\nu_0)) \quad \forall z \in \Phi(\mu,\nu) \cap \mathbb{B}(z^0;\varepsilon), \tag{2.3}$$

$$d(z,\Phi(\mu,\nu)) \le L\alpha((\mu,\nu),(\mu_0,\nu_0)) \quad \forall z \in \Phi(\mu_0,\nu_0) \cap \mathbb{B}(z^0;\varepsilon). \tag{2.4}$$

4. *f is locally Lipschitz.*

Then, for any bounded and open set Q containing the original solution set, the following stability properties hold true:

- *$\exists \delta' > 0 :\ \Psi_Q(\mu,\nu) \neq \emptyset$ for all $(\mu,\nu) \in \mathbb{B}((\mu_0,\nu_0);\delta')$.*
- *Ψ_Q is upper semicontinuous at (μ_0,ν_0) in the sense of Berge; i.e., for all open $V \supseteq \Psi(\mu_0,\nu_0) = \Psi_Q(\mu_0,\nu_0)$ there exists some $\delta_V > 0$ such that*

$$\Psi_Q(\mu,\nu) \subseteq V \quad \forall (\mu,\nu) \in \mathbb{B}((\mu_0,\nu_0);\delta_V).$$

- *φ_Q is continuous at (μ_0,ν_0) and satisfies the following Lipschitz-like estimate for some constants $\delta^*, L^* > 0$:*

$$|\varphi_Q(\mu,\nu) - \varphi_Q(\mu_0,\nu_0)| \le L^*\alpha((\mu,\nu),(\mu_0,\nu_0)) \quad \forall (\mu,\nu) \in \mathbb{B}((\mu_0,\nu_0);\delta^*).$$

We note that the first two assertions of the theorem already follow from [19, Theorem 4.3]. In the following we want to provide verifiable conditions for the assumptions of Theorem 2.1. As far as assumption 1 is concerned, it is of a purely technical nature and may be difficult to verify in the general setting. If, however, the abstract part Z of the constraint set in (2.1) happens to be compact, as is the case in many applied problems, then, of course, the boundedness assumption 1 in Theorem 2.1 is trivially satisfied. In this situation, one can even drop the localizations φ_Q and Ψ_Q in the statement of Theorem 2.1 and formulate the corresponding conclusions for the global optimal value function φ and the global solution set mapping Ψ. Indeed, as one may choose Q in Theorem 2.1 by compactness of Z such that $Q \supseteq Z \supseteq \Psi(\mu_0,\nu_0)$, it follows that

$$\Psi_Q(\mu,\nu) = \Psi(\mu,\nu)\ (\subset Z \subset Q) \quad \text{and} \quad \varphi_Q(\mu,\nu) = \varphi(\mu,\nu) \quad \forall (\mu,\nu).$$

Passing to assumption 2 in Theorem 2.1, this is generally satisfied under the data assumptions made for problem (1.3). To show this, we first adapt a result of [22].

LEMMA 2.2. *Assume that a multifunction $S : \mathbb{R}^n \rightrightarrows \mathbb{R}^s$ has a closed graph. Let $\bar{x} \in \mathbb{R}^n$ be such that $S(\bar{x}) \neq \emptyset$. Then for every nonnegative regular measure μ on $\mathbb{R}^s$ and for every $\varepsilon > 0$ there exists $\delta > 0$ such that*

$$\mu\big(S(x)\big) \le \mu\big(S(\bar{x})\big) + \varepsilon, \ \textit{whenever } \|x - \bar{x}\| \le \delta. \tag{2.5}$$

Proof. By the closedness of the graph,

$$S(\bar{x}) = \bigcap_{\delta>0} \mathrm{cl}\Big(\bigcup_{\|x-\bar{x}\|\le\delta} S(x)\Big).$$

Therefore, for every regular measure μ,

$$\mu\big(S(\bar{x})\big) = \inf_{\delta>0} \mu\Big(\mathrm{cl}\Big(\bigcup_{\|x-\bar{x}\|\le\delta} S(x)\Big)\Big).$$

Consequently, for every $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\mu\big(S(\bar{x})\big) + \varepsilon \ge \mu\Big(\mathrm{cl}\Big(\bigcup_{\|x-\bar{x}\|\le\delta} S(x)\Big)\Big).$$

This implies the result. □

THEOREM 2.3. *The graph of the feasible set mapping Φ is closed.*

Proof. Consider a sequence (μ^n, ν^n, z^n) of the elements of the graph, which is convergent to some $(\bar{\mu}, \bar{\nu}, \bar{z})$ in the space $\mathscr{P}(\mathbb{R}^s) \times \mathscr{P}(\mathbb{R}) \times \mathbb{R}^n$. Since $z^n \in \Phi(\mu^n, \nu^n)$, then $z^n \in Z$ and

$$\mu^n(H(z^n, \eta)) - \nu^n([\eta, \infty)) \ge 0 \quad \forall \eta \in [a, b]. \tag{2.6}$$

As Z is closed, $\bar{z} \in Z$. By the definition of $\alpha_1(\cdot,\cdot)$, it follows that

$$\nu^n([\eta, \infty)) \to \bar{\nu}([\eta, \infty)) \quad \forall\, \eta \in [a, b]. \tag{2.7}$$

Let us consider the first term in (2.6). For a fixed $\eta \in [a, b]$ we have the inequality

$$\begin{aligned} &\mu^n(H(z^n, \eta)) - \bar{\mu}(H(\bar{z}, \eta)) \\ &= \big[\mu^n(H(z^n, \eta)) - \bar{\mu}(H(z^n, \eta))\big] + \big[\bar{\mu}(H(z^n, \eta)) - \bar{\mu}(H(\bar{z}, \eta))\big] \\ &\le \alpha_{\mathrm{B}}\,(\mu^n, \bar{\mu}) + \big[\bar{\mu}(H(z^n, \eta)) - \bar{\mu}(H(\bar{z}, \eta))\big]. \end{aligned} \tag{2.8}$$

By assumption, $\alpha_{\mathrm{B}}\,(\mu^n, \bar{\mu}) \to 0$, and we can focus on the term in brackets. By the continuity of g, the multifunction $H(\cdot, \eta)$ has a closed graph. We now apply Lemma 2.2 to conclude that for every $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\bar{\mu}\big(H(z, \eta)\big) \le \bar{\mu}\big(H(\bar{z}, \eta)\big) + \varepsilon, \text{ whenever } \|z - \bar{z}\| \le \delta.$$

For all sufficiently large n one has $\|z^n - \bar{z}\| \le \delta$ and therefore

$$\bar{\mu}\big(H(z^n, \eta)\big) \le \bar{\mu}\big(H(\bar{z}, \eta)\big) + \varepsilon.$$

Passing to the limit with $n \to \infty$ and noting that $\varepsilon > 0$ was arbitrary, we obtain

$$\limsup_{n\to\infty} \bar{\mu}\big(H(z^n, \eta)\big) \le \bar{\mu}\big(H(\bar{z}, \eta)\big). \tag{2.9}$$

Combining relations (2.8) and (2.9), we conclude that

$$\limsup_{n\to\infty} \mu^n(H(z^n, \eta)) \le \bar{\mu}(H(\bar{z}, \eta)).$$

Using this in (2.6), with a view to (2.7), we obtain

$$\begin{aligned} \bar{\mu}(H(\bar{z}, \eta)) - \bar{\nu}(\eta, \infty)) &\ge \limsup_{n\to\infty} \mu^n(H(z^n, \eta)) - \lim_{n\to\infty} \nu^n([\eta, \infty)) \\ &= \limsup_{n\to\infty} \big[\mu^n(H(z^n, \eta)) - \nu^n([\eta, \infty))\big] \ge 0. \end{aligned}$$

Since η was arbitrary, we obtain the relation

$$\bar{\mu}(H(\bar{z},\eta)) - \bar{\nu}(\eta,\infty)) \geq 0 \quad \forall\, \eta \in [a,b].$$

This amounts to $\bar{z} \in \Phi(\bar{\mu},\bar{\nu})$, as desired. □

Remark 1. Let us observe that we did not use the compactness of the set $[a,b]$ in the proof, and therefore Theorem 2.3 holds true for the dominance relation enforced on the whole real line.

The verification of assumption 3 in Theorem 2.1 is less direct and will be based on an appropriate constraint qualification for problem (2.1) at the original parameter (μ_0,ν_0). To formulate this constraint qualification, we assume the following differential uniform dominance condition introduced in [5].

DEFINITION 2.4. *Problem* (2.1) *for* $\mu = \mu_0$ *and* $\nu = \nu_0$ *satisfies the* differential uniform dominance condition *at the point* $z^0 \in Z$ *if*

(i) $\mu_0(H(z,\eta))$ *is continuous with respect to* η *in* $[a,b]$*, differentiable with respect to* z *in a neighborhood of* z^0 *for all* $\eta \in [a,b]$*, and its derivative is jointly continuous with respect to both arguments;*

(ii) $\nu_0([\cdot,\infty))$ *is continuous;*

(iii) *there exists* $z^1 \in Z$ *such that*

$$\min_{a\leq\eta\leq b} \Big\{\mu_0\big(H(z^0,\eta)\big) + \nabla_z\mu_0\big(H(z^0,\eta)\big)(z^1 - z^0) - \nu_0\big([\eta,\infty)\big)\Big\} > 0.$$

The differentiability assumptions on $\mu_0(H(\cdot,\eta))$ can be guaranteed by assuming continuous differentiability of the function g with respect to both arguments, the existence of the probability density of the random vector V, and by mild regularity conditions (see [9]). Then

$$\nabla_z\mu_0\big(H(z,\eta)\big) = \int_{\partial H(z,\eta)} \frac{\varphi(v)}{\|\nabla_v g(z,v)\|}\nabla_z g(z,v)\,\lambda(dv),$$

where $\partial H(z,\eta)$ is the surface of the set $H(z,\eta)$ and λ is the surface Lebesgue measure. The regularity conditions mentioned require that the gradient $\nabla_v g(z,v)$ be nonzero and that the integrand above be uniformly bounded (in a neighborhood of z) by an integrable function.

For example, if $g(z,V) = \langle z,V\rangle$ and V has a nondegenerate multivariate normal distribution $\mathcal{N}(\bar{v},\Sigma)$, then

$$\mu_0(H(z,\eta)) = 1 - \Phi\Big(\frac{\eta - \langle z,\bar{v}\rangle}{\sqrt{\langle z,\Sigma z\rangle}}\Big),$$

where $\Phi(\cdot)$ is the distribution function of the standard normal variable. In this case condition (i) of Definition 2.4 is satisfied at every $z \neq 0$.

The differential uniform dominance condition has substantial consequences. Let $\mathscr{C}$ be the Banach space of continuous functions on $[a,b]$. Consider the mapping $\Gamma : \mathbb{R}^n \to \mathscr{C}$ defined as

$$\Gamma(z)(\eta) = \mu_0(H(z,\eta)) - \nu_0([\eta,\infty)), \quad \eta \in [a,b],$$

where $\nu_0([\cdot,\infty)) \in \mathscr{C}$. Denote by K the nonnegative cone in $\mathscr{C}$.

LEMMA 2.5. *Assume that* $\mu_0(H(z,\eta))$ *is continuously differentiable with respect to* z *in a neighborhood of* $z^0 \in Z$ *and for all* $\eta \in [a,b]$*,* $\mu_0(H(z,\cdot))$ *is continuous in*

$[a,b]$, and $\Gamma(z^0) \in K$. The differential uniform dominance condition is satisfied at z^0 if and only if the multifunction

$$z \mapsto \begin{cases} \Gamma(z) - K & \text{if } z \in Z, \\ \emptyset & \text{otherwise} \end{cases} \tag{2.10}$$

is metrically regular at $(z^0, 0)$.

Proof. We observe that the differential uniform dominance condition is equivalent to Robinson's constraint qualification condition (see [18])

$$0 \in \operatorname{int}\Big\{\Gamma(z^0) + \nabla_z \Gamma(z^0)(Z - z^0) - K\Big\}. \tag{2.11}$$

Indeed, it is easy to see that the uniform dominance condition implies Robinson's condition. On the other hand, if Robinson's condition holds true, then there exists $\varepsilon > 0$ such that the function identically equal to ε is an element of the set on the right-hand side of (2.11). Then we can find z^1 such that

$$\Gamma(z^0)(\eta) + \big[\nabla_z \Gamma(z^0)(\eta)\big](z^1 - z^0) \geq \varepsilon \quad \forall\, \eta \in [a,b].$$

Consequently, the uniform dominance condition is satisfied. On the other hand, Robinson's constraint qualification at z^0 is equivalent to the metric regularity of (2.10) at $(z^0, 0)$ (see [2]). □

The next proposition shows that the verification of assumption 3 in Theorem 2.1 can be reduced to the differential uniform dominance condition.

PROPOSITION 2.6. *Let the differential uniform dominance condition be satisfied at some $z^0 \in \Phi(\mu_0, \nu_0)$. Then relations* (2.3) *and* (2.4) *of Theorem* 2.1 *hold true at z^0.*

Proof. We introduce the multifunction $M : \mathscr{C} \rightrightarrows \mathbb{R}^n$ as the following parameter dependent constraint set mapping:

$$M(w) := \big\{z \in Z : \mu_0(H(z,\eta)) - w(\eta) \geq 0 \ \forall\, \eta \in [a,b]\big\}.$$

(The relation between M and Φ is given by $\Phi(\mu_0, \nu) = M\big(\nu([\cdot, \infty))\big)$ for all continuous distributions $\nu \in \mathscr{P}(\mathbb{R})$.) Define $w_0(\cdot) = \nu_0([\cdot, \infty))$. By assumption, $w_0 \in \mathscr{C}$.

By Lemma 2.5, the differential uniform dominance condition is equivalent to metric regularity of (2.10) at $(z^0, 0)$, which, upon passing to the inverse multifunction, is equivalent to the pseudo-Lipschitz property of M at (w^0, z^0) (see, e.g., [12, Lemma 1.12] and [20, Theorem 9.43]). Accordingly, there exist $\tilde{\varepsilon} > 0$ and $\tilde{L} > 0$ such that

$$d(z, M(w_2)) \leq \tilde{L} d(w_1, w_2) \quad \forall z \in M(w_1) \cap \mathbb{B}(z^0; \tilde{\varepsilon}) \ \forall w_1, w_2 \in \mathbb{B}(w_0; \tilde{\varepsilon}), \tag{2.12}$$

where the last ball is taken in the metric of $\mathscr{C}$. First, we verify the following chain of inclusions for all $(\mu, \nu) \in \mathscr{P}(\mathbb{R}^s) \times \mathscr{P}(\mathbb{R})$:

$$M(w_0 + 2\alpha((\mu,\nu),(\mu_0,\nu_0)) \cdot \mathbb{1}) \subseteq \Phi(\mu,\nu) \subseteq M(w_0 - 2\alpha((\mu,\nu),(\mu_0,\nu_0)) \cdot \mathbb{1}), \tag{2.13}$$

where $\mathbb{1}$ is the function on $[a,b]$ taking the constant value 1. Note that M is applied to continuous functions as required. Now, if

$$z \in M(w_0 + 2\alpha_{\mathrm{B}}\,((\mu,\nu),(\mu_0,\nu_0)) \cdot \mathbb{1}),$$

then $z \in Z$ and, by definition of α,

$$\begin{aligned} 0 &\le \mu_0(H(z,\eta)) - (w_0(\eta) + 2\alpha((\mu,\nu),(\mu_0,\nu_0))) \\ &= \mu_0(H(z,\eta)) - \nu_0([\eta,\infty)) - 2\alpha((\mu,\nu),(\mu_0,\nu_0))) \\ &\le \mu(H(z,\eta)) - \nu([\eta,\infty)) \quad \forall \eta \in [a,b]\,. \end{aligned}$$

This establishes the first inclusion of (2.13), and the second one is completely analogous.

In order to check (2.3), let $(\mu,\nu) \in \mathbb{B}((\mu_0,\nu_0);\tilde{\varepsilon}/2)$ and $z \in \Phi(\mu,\nu) \cap \mathbb{B}(z^0;\tilde{\varepsilon}/2)$ be arbitrary. Define $w_1 \in \mathscr{C}$ by $w_1 := w_0 - 2\alpha((\mu,\nu),(\mu_0,\nu_0)) \cdot \mathbb{1}$. Then the second inclusion of (2.13) entails that $z \in M(w_1)$. Furthermore,

$$d(w_1,w_0) = 2\alpha((\mu,\nu),(\mu_0,\nu_0)) \le \tilde{\varepsilon}.$$

Consequently, we may apply (2.12) to w_1 and to $w_2 := w_0 \in \mathscr{C}$:

$$d(z,\Phi(\mu_0,\nu_0)) = d(z,M(w_0)) \le \tilde{L}d(w_1,w_0) = 2\tilde{L}\alpha((\mu,\nu),(\mu_0,\nu_0)).$$

Therefore, (2.3) holds true with $L := 2\tilde{L}$ and $\varepsilon := \tilde{\varepsilon}/2$. As for (2.4), take arbitrary $(\mu,\nu) \in \mathbb{B}((\mu_0,\nu_0);\tilde{\varepsilon}/2)$ and $z \in \Phi(\mu_0,\nu_0) \cap \mathbb{B}(z^0;\tilde{\varepsilon}/2)$. Define $w_2 \in \mathscr{C}$ by $w_2 := w_0 + 2\alpha((\mu,\nu),(\mu_0,\nu_0)) \cdot \mathbb{1}$. Then

$$d(w_2,w_0) = 2\alpha((\mu,\nu),(\mu_0,\nu_0)) \le \tilde{\varepsilon},$$

and we may apply (2.12) to $w_1 := w_0$ and to w_2. Further taking into account the first inclusion of (2.13), one arrives at

$$d(z,\Phi(\mu,\nu)) \le d(z,M(w_2) \le \tilde{L}d(w_0,w_2) = 2\tilde{L}\alpha((\mu,\nu),(\mu_0,\nu_0)),$$

which is (2.4) with the same values $L := 2\tilde{L}$ and $\varepsilon := \tilde{\varepsilon}/2$ as for (2.3). □

3. Sensitivity of the optimal value.

3.1. Optimality conditions. In order to analyze the sensitivity of the optimal value function, we need to briefly recall optimality conditions for problem (1.3). From now on we assume that f is continuously differentiable.

We define the set $\mathscr{U}([a,b])$ of functions $u(\cdot)$ satisfying the following conditions:

$$\begin{aligned} &u(\cdot) \text{ is nondecreasing and right continuous;} \\ &u(t) = 0 \;\forall\; t \le a; \\ &u(t) = u(b) \;\forall\, t \ge b. \end{aligned}$$

It is evident that $\mathscr{U}([a,b])$ is a convex cone. The slight difference from the definition of the set $\mathscr{U}$ introduced in [5] is due to the fact that we formulate the stochastic dominance constraint in (1.3) via the $\ge$ inequality.

We introduce the functional $L : \mathbb{R}^n \times \mathscr{U}([a,b]) \times \mathscr{P}(\mathbb{R}^s) \times \mathscr{P}(\mathbb{R}) \to \mathbb{R}$ associated with problem (1.3):

$$L(z,u;\mu,\nu) := f(z) - \int u(g(z,v))\,\mu(dv) + \int u(y)\,\nu(dy). \tag{3.1}$$

As shown in [5], the functional L plays a similar role to that of a Lagrangian of the problem.

THEOREM 3.1. *Assume that the differential uniform dominance condition is satisfied at a local minimum $\hat{z}$ of problem* (1.3). *Then there exists a function $\hat{u} \in \mathscr{U}([a,b])$ such that*

$$-\nabla_z L(\hat{z}, \hat{u}; \mu_0, \nu_0) \in N_Z(\hat{z}), \tag{3.2}$$

$$\int \hat{u}(g(\hat{z}, v))\, \mu_0(dv) = \int \hat{u}(y)\, \nu_0(dy). \tag{3.3}$$

The proof follows the same line of argument as the proof in [5] and is omitted here. It uses the correspondence between a nonnegative measure λ on $[a,b]$ and a function $u \in \mathscr{U}([a,b])$:

$$u(\eta) = \lambda([a,\eta]), \quad \eta \in [a,b]. \tag{3.4}$$

Remark 2. The set $\hat{U}(\hat{z})$ of functions in $\mathscr{U}([a,b])$ satisfying (3.2)–(3.3) for the local minimum $\hat{z}$ is convex, bounded, and weakly* closed in the following sense: if a sequence of functions $u^k \in \hat{U}(\hat{z})$ and $u \in \mathscr{U}([a,b])$ are such that

$$\lim_{k\to\infty} \int_a^b c(\eta)\, du^k(\eta) = \int_a^b c(\eta)\, du(\eta) \quad \forall\, c \in \mathscr{C},$$

then $u \in \hat{U}(\hat{z})$. This follows from [1, Theorem 3.6] and the application of (3.4).

If the function $g(\cdot,\cdot)$ is quasi-concave and μ has an r-concave probability density function, with $r \geq -1/s$, then the feasible set of problem (1.3) is convex (see [16]). Therefore we can formulate the following sufficient conditions of optimality, as in [5].

THEOREM 3.2. *Assume that a point $\hat{z}$ is feasible for problem* (1.3). *Suppose that there exists a function $\hat{u} \in \mathscr{U}([a,b])$ such that conditions* (3.2)–(3.3) *are satisfied. If the function f is convex, the function $g(\cdot,\cdot)$ is quasi-concave, and V has an r-concave probability density function, with $r \geq -1/s$, then $\hat{z}$ is an optimal solution of problem* (1.3).

Let us observe that under the assumptions of Theorem 3.2 the functional (3.1) is, in general, not a quasi-convex function of z.

3.2. Upper bound. Consider the measures

$$\mu_t = \mu_0 + t\gamma,$$
$$\nu_t = \nu_0 + t\sigma,$$

where γ and σ are regular signed measures on $\mathbb{R}^s$ and $\mathbb{R}$, respectively, and $t > 0$. We shall bound the optimal value $\varphi(\mu_t, \nu_t)$ of the perturbed problem

$$\begin{aligned} &\min\ f(z) \\ &\text{s.t. } \mu_t(H(z,\eta)) - \nu_t([\eta,\infty)) \geq 0 \quad \forall \eta \in [a,b], \\ &\qquad z \in Z. \end{aligned} \tag{3.5}$$

Our objective is to develop bounds for the limit of the quotients $\big[\varphi(\mu_t,\nu_t) - \varphi(\mu_0,\nu_0)\big]/t$, when $t \downarrow 0$.

THEOREM 3.3. *Let $\hat{Z}$ be the set of optimal solutions of problem* (1.3). *Assume the following conditions:*

(i) *The differential uniform dominance condition is satisfied at each point $\hat{z} \in \hat{Z}$.*

(ii) $\gamma(H(z,\eta))$ *is continuous with respect to both arguments at* $(\hat{z},\eta)$ *for all* $\eta \in [a,b]$*, is differentiable with respect to* z *in a neighborhood of each* $\hat{z} \in \hat{Z}$ *for every value of* $\eta \in [a,b]$*, and its derivative is jointly continuous with respect to both arguments.*

(iii) $\sigma([\eta,\infty))$ *is a continuous function of* η.

Then

$$\begin{aligned}\limsup_{t\downarrow 0} \frac{1}{t}[\varphi(\mu_t,\nu_t) - \varphi(\mu_0,\nu_0)] \\ \leq \inf_{\hat{z}\in\hat{Z}} \sup_{\hat{u}\in\hat{U}(\hat{z})} \Big\{ \int \hat{u}(g(\hat{z},v))\,\gamma(dv) + \int \hat{u}(y)\,\sigma(dy)\Big\},\end{aligned} \tag{3.6}$$

where $\hat{U}(\hat{z})$ *is the set of functions in* $\mathscr{U}([a,b])$ *satisfying* (3.2)–(3.3) *at the minimum* $\hat{z}$.

Proof. Our result is close in spirit to that of [1, Proposition 4.22], but we work with weaker assumptions by exploiting the structure of the problem.

Fix $\hat{z} \in \hat{Z}$. We shall construct feasible points of the perturbed problem of the form

$$\tilde{z}_t = \hat{z} + th + o(t). \tag{3.7}$$

Define the set

$$\mathscr{A} = \big\{\eta \in [a,b] : \mu_0(H(\hat{z},\eta)) = \nu_0([\eta,\infty))\big\},$$

and let $T_Z(\hat{z})$ denote the tangent cone to Z at $\hat{z}$.

We assume that the direction h in (3.7) is an element of the tangent cone $T_Z(\hat{z})$ and satisfies the infinite system of linear inequalities:

$$\langle \nabla_z \mu_0(H(\hat{z},\eta)), h\rangle + \gamma(H(\hat{z},\eta)) - \sigma([\eta,\infty)) \geq 0 \quad \forall\, \eta \in \mathscr{A}. \tag{3.8}$$

It follows from the uniform dominance condition that there exists $\varepsilon > 0$ such that

$$\langle \nabla_z \mu_0(H(\hat{z},\eta)), z^1 - \hat{z}\rangle > \varepsilon$$

for all $\eta \in \mathscr{A}$. Therefore inequalities (3.8) can be satisfied by choosing $h = \tau(z^1 - \hat{z})$ with a sufficiently large τ.

Let $z_t = \hat{z} + th$. The uniform dominance condition implies that

$$\begin{aligned}\mu_t(H(z_t,\eta)) &= \mu_0(H(z_t,\eta)) + t\gamma(H(z_t,\eta)) \\ &= \mu_0(H(\hat{z},\eta)) + t\langle \nabla_z \mu_0(H(\hat{z},\eta)), h\rangle + t\gamma(H(z_t,\eta)) + o(t,\eta),\end{aligned} \tag{3.9}$$

where $o(t,\eta)/t \to 0$ as $t \to 0$, uniformly over $\eta \in [a,b]$.

We shall estimate the term $\gamma(H(z_t,\eta))$ from below. Choose any $\hat{\eta} \in [a,b]$. By the continuity of $\gamma(H(z,\eta))$ around the point $(\hat{z},\hat{\eta})$, for every $\varepsilon > 0$ there exists $\delta(\varepsilon,\hat{\eta}) > 0$ such that

$$\gamma(H(z,\eta)) \geq \gamma(H(\hat{z},\hat{\eta})) - \varepsilon \tag{3.10}$$

for all (z,η) such that $\|z - \hat{z}\| \leq \delta(\varepsilon,\hat{\eta})$ and $|\eta - \hat{\eta}| \leq \delta(\varepsilon,\hat{\eta})$. For each ε the intervals $|\eta - \hat{\eta}| \leq \delta(\varepsilon,\hat{\eta})$, where $\hat{\eta}$ runs through $[a,b]$, cover $[a,b]$. Choosing a finite subcovering, we conclude that there exists $\delta(\varepsilon) > 0$ such that (3.10) holds true for all z satisfying $\|z - \hat{z}\| \leq \delta(\varepsilon)$ and for all $\eta \in [a,b]$.

Define $r(t) = \inf\{\varepsilon > 0 : \delta(\varepsilon) \geq t\|h\|\}$. Observe that $r(t) \to 0$ as $t \downarrow 0$. It follows from (3.10) that

$$\gamma(H(z_t, \eta)) \geq \gamma(H(\hat{z}, \eta)) - r(t).$$

Substituting this estimate into (3.9), we obtain

$$\mu_t(H(z_t, \eta)) \geq \mu_0(H(\hat{z}, \eta)) + t\langle \nabla_z \mu_0(H(\hat{z}, \eta)), h\rangle + t\gamma(H(\hat{z}, \eta)) + o(t, \eta) - tr(t).$$

Using condition (3.8) and the feasibility of $\hat{z}$, we conclude that

$$\begin{aligned} \mu_t(H(z_t, \eta)) - \nu_t([\eta, \infty)) &= \big[\mu_0(H(\hat{z}, \eta)) - \nu_0([\eta, \infty))\big] \\ &\quad + t\big[\langle \nabla_z \mu_0(H(\hat{z}, \eta)), h\rangle + \gamma(H(\hat{z}, \eta)) - \sigma([\eta, \infty))\big] \\ &\quad + o(t, \eta) - tr(t) \\ &\geq o(t, \eta) - tr(t) \quad \forall\, \eta \in [a, b]. \end{aligned} \tag{3.11}$$

Consequently, the point z_t may violate the constraints of the perturbed problem only by quantities which are infinitely smaller than t. Define the mapping $\Gamma : \mathbb{R}^n \times \mathbb{R} \to \mathscr{C}$ as follows:

$$\Gamma(z, t)(\eta) = \mu_t(H(z, \eta)) - \nu_t([\eta, \infty)), \quad \eta \in [a, b].$$

The system

$$\begin{aligned} \Gamma(z, t) &\in K, \\ z &\in Z, \end{aligned}$$

is stable about $(\hat{z}, 0)$ (see, e.g., [1, Theorem 2.87]). Therefore, for all sufficiently small $t > 0$, we can slightly modify z_t to get a point $\tilde{z}_t$ such that

$$\begin{aligned} \Gamma(\tilde{z}_t, t) &\in K, \\ \tilde{z}_t &\in Z, \\ \|\tilde{z}_t - z_t\| &\leq C\big[\operatorname{dist}(\Gamma(z_t, t), K) + \operatorname{dist}(z_t, Z)\big], \end{aligned}$$

where C is some constant. Using (3.11) and the fact that h is tangent to Z, we obtain that

$$\lim_{t \downarrow 0} \frac{1}{t}\big(\tilde{z}_t - \hat{z}\big) = h.$$

As $\tilde{z}_t$ is feasible,

$$\varphi(\mu_t, \nu_t) \leq f(\tilde{z}_t).$$

Subtracting $\varphi(\mu_0, \nu_0)$, dividing by t, and passing to the limit, we obtain

$$\limsup_{t \downarrow 0} \frac{1}{t}[\varphi(\mu_t, \nu_t) - \varphi(\mu_0, \nu_0)] \leq \limsup_{t \downarrow 0} \frac{1}{t}[f(\tilde{z}_t) - f(\hat{z})] = \langle \nabla f(\hat{z}), h\rangle. \tag{3.12}$$

It follows that the limit on the left-hand side of (3.12) is bounded from above by the optimal value of the problem

$$\begin{aligned} &\min\ \langle \nabla f(\hat{z}), h\rangle \\ &\text{s.t. } \langle \nabla_z \mu_0(H(\hat{z}, \eta)), h\rangle \geq -\gamma(H(\hat{z}, \eta)) + \sigma([\eta, \infty)) \quad \forall\, \eta \in \mathscr{A}, \\ &\quad h \in T_Z(\hat{z}). \end{aligned} \tag{3.13}$$

The optimal value of the linear-conic problem (3.13) is equal to the optimal value of the following dual problem (see, e.g., [1, Theorem 5.106]):

$$\begin{aligned} &\max_{\lambda} \int_a^b \big[-\gamma(H(\hat z,\eta)) + \sigma([\eta,\infty))\big]\,\lambda(d\eta) \\ &\text{s.t. } -\nabla f(\hat z) - \int_a^b \nabla_z \mu_0(H(\hat z,\eta))\,\lambda(d\eta) \in N_Z(\hat z), \\ &\quad \lambda \ge 0. \end{aligned} \tag{3.14}$$

Here λ is a regular measure on $\mathscr{A}$. Moreover, it is sufficient to consider atomic measures λ with at most $n+1$ atoms.

Extending λ to $[a,b]$, associating with it a function $u(\cdot)=\lambda([a,\cdot])$, and changing the order of integration, we obtain the identity

$$\begin{aligned} \int_a^b \gamma(H(\hat z,\eta))\,\lambda(d\eta) &= \int_a^b \int_{v\in H(\hat z,\eta)} \gamma(dv)\,\lambda(d\eta) = \int_a^b \int_{\{v:g(\hat z,v)\ge\eta\}} \gamma(dv)\,\lambda(d\eta) \\ &= \int\int_a^{g(\hat z,v)} \lambda(d\eta)\,\gamma(dv) = \int u(g(\hat z,v))\,\gamma(dv). \end{aligned} \tag{3.15}$$

In a similar way we transform other integrals in (3.14) to obtain the following form of the dual problem:

$$\begin{aligned} &\max_{u(\cdot)} -\int u(g(\hat z,v))\,\gamma(dv) + \int u(y)\,\sigma(dy) \\ &\text{s.t. } -\nabla f(\hat z) - \nabla_z \int u(g(\hat z,v))\,\mu_0(dv) \in N_Z(\hat z), \\ &\quad u(\cdot) \in \mathscr{U}([a,b]), \\ &\quad u(\cdot) \text{ satisfies (3.3)}. \end{aligned} \tag{3.16}$$

We observe that the feasible set of this problem is the set $\hat U$ given by (3.2)–(3.3). Now we continue the estimate (3.12) as follows:

$$\limsup_{t\downarrow 0} \frac{1}{t}[\varphi(\mu_t,\nu_t)-\varphi(\mu_0,\nu_0)] \le \sup_{\hat u\in\hat U(\hat z)} \Big\{ -\int \hat u(g(\hat z,v))\,\gamma(dv) + \int \hat u(y)\,\sigma(dy)\Big\}.$$

As $\hat z\in\hat Z$ was arbitrary, we conclude that

$$\limsup_{t\downarrow 0} \frac{1}{t}[\varphi(\mu_t,\nu_t)-\varphi(\mu_0,\nu_0)] \le \inf_{\hat z\in\hat Z} \sup_{\hat u\in\hat U(\hat z)} \Big\{ -\int \hat u(g(\hat z,v))\,\gamma(dv) + \int \hat u(y)\,\sigma(dy)\Big\},$$

which was what we set out to prove. □

As discussed in the proof, it is sufficient to consider the supremum over piecewise constant functions $\hat u\in\hat U$ having at most $n+1$ jumps.

COROLLARY 3.4. *Suppose that $\mu_1=\mu_0+\gamma$ is a nonnegative measure and let $\nu_1=\nu_0+\sigma$. Then*

$$\limsup_{t\downarrow 0} \frac{1}{t}[\varphi(\mu_t,\nu_t)-\varphi(\mu_0,\nu_0)] \le \inf_{\hat z\in\hat Z} \sup_{\hat u\in\hat U(\hat z)} \int \hat u(y)\,\nu_1(dy).$$

Proof. We can rewrite the estimate (3.6) as follows:

$$\limsup_{t\downarrow 0}\frac{1}{t}[\varphi(\mu_t,\nu_t)-\varphi(\mu_0,\nu_0)]\le \inf_{\hat z\in\hat Z}\sup_{\hat u\in\hat U(\hat z)}\Big\{\int \hat u(g(\hat z,v))\,\mu_0(dv) - \int \hat u(g(\hat z,v))\,\mu_1(dv)+\int \hat u(y)\,\nu_1(dy)-\int \hat u(y)\,\nu_0(dy)\Big\}.$$

As the function $\hat u(\cdot)$ is nonnegative, we can skip the second term on the right-hand side. Using the complementarity condition (3.3), we get the required inequality. ∎

3.3. Lower bound. Let us start from the following observation.

LEMMA 3.5. *Consider any measures $\mu\in\mathscr{P}(\mathbb{R}^s)$ and $\nu\in\mathscr{P}(\mathbb{R})$ and a point $z\in Z$ such that*

$$\mu(H(z,\eta))\ge\nu([\eta,\infty)),\quad \eta\in[a,b]. \tag{3.17}$$

Then for every $u\in\mathscr{U}([a,b])$ we have

$$\int u(g(z,v))\,\mu(dv)\ge\int u(y)\,\nu(dy).$$

Proof. For a function $u\in\mathscr{U}([a,b])$ we define a nonnegative measure λ on $[a,b]$ by the relation $u(\cdot)=\lambda([a,\cdot])$. Integrating the inequalities (3.17), changing the order of integration as in (3.15), we obtain the postulated inequality. ∎

Suppose that $u\in\mathscr{U}([a,b])$. Employing Lemma 3.5, we obtain

$$\varphi(\mu,\nu)\ge\inf_{z\in Z}\Big\{f(z)-\int u(g(z,v))\,\mu(dv)\Big\}+\int u(y)\,\nu(dy).$$

We get the general dual lower bound

$$\varphi(\mu,\nu)\ge\sup_{u\in\mathrm{U}([a,b])}\inf_{z\in Z}\Big\{f(z)-\int u(g(z,v))\,\mu(dv)+\int u(y)\,\nu(dy)\Big\}.$$

In order to obtain tighter bounds we consider the perturbations in directions

$$\mu_t=\mu_0+t\gamma,$$
$$\nu_t=\nu_0+t\sigma.$$

We shall develop lower bounds for the differential quotients $\big[\varphi(\mu_t,\nu_t)-\varphi(\mu_0,\nu_0)\big]/t$ when $t\downarrow 0$. Our result is similar to the standard approach employed in [1, Theorem 4.24]. However, it is unrealistic to assume that the Lagrangian is convex (even under the assumptions of Theorem 3.2), and that is why we need Lipschitz stability of optimal solutions.

THEOREM 3.6. *Assume that $\hat z$ is the unique optimal solution of problem* (1.3) *and that the differential uniform dominance condition is satisfied at $\hat z$. Furthermore, assume that the perturbed problems* (3.5) *have solutions z_t such that $\|z_t-\hat z\|\le Lt$ with some constant L. Let $\hat U$ be the set of functions $\hat u(\cdot)$ satisfying the optimality conditions* (3.2)–(3.3). *Then*

$$\liminf_{t\to 0}\frac{1}{t}\big[\varphi(\mu_t,\nu_t)-\varphi(\mu_0,\nu_0)\big] \ge\sup_{\hat u\in\hat U}\Big\{-\int\hat u\big(g(\hat z,v)\big)\,\gamma(dv)+\int\hat u(y)\,\sigma(dy)\Big\}. \tag{3.18}$$

Proof. Consider problem (2.1) and its Lagrangian

$$\Lambda(z,\lambda;\mu,\nu) = f(z) - \int_a^b \big[\mu(H(z,\eta)) - \nu([\eta,\infty))\big]\,\lambda(d\eta),$$

where λ is a nonnegative regular measure on $[a,b]$. Fix $\mu = \mu_0$ and $\nu = \nu_0$. As in [5], owing to the differential uniform dominance condition at $\hat{z}$, there exists a measure $\hat{\lambda} \geq 0$ such that

$$\langle \nabla_z \Lambda(\hat{z},\hat{\lambda};\mu_0,\nu_0), z - \hat{z}\rangle \geq 0 \quad \forall\, z \in Z$$

and

$$\int_a^b \big[\mu_0(H(\hat{z},\eta)) - \nu_0([\eta,\infty))\big]\,\hat{\lambda}(d\eta) = 0.$$

Using the nonnegativity of $\hat{\lambda}$ and the complementarity condition, we can write the chain of inequalities

$$\begin{aligned}
&\varphi(\mu_t,\nu_t) - \varphi(\mu_0,\nu_0) \geq f(z_t) - \int_a^b \big[\mu_t(H(z_t,\eta)) - \nu_t([\eta,\infty))\big]\,\hat{\lambda}(d\eta) - f(\hat{z})\\
&\geq f(z_t) - \int_a^b \big[\mu_t(H(z_t,\eta)) - \nu_t([\eta,\infty))\big]\,\hat{\lambda}(d\eta)\\
&\quad - f(\hat{z}) + \int_a^b \big[\mu_0(H(\hat{z},\eta)) - \nu_0([\eta,\infty))\big]\,\hat{\lambda}(d\eta)\\
&= \Lambda(z_t,\hat{\lambda};\mu_0,\nu_0) - \Lambda(\hat{z},\hat{\lambda};\mu_0,\nu_0) - t\int_a^b \big[\gamma(H(z_t,\eta)) - \sigma([\eta,\infty))\big]\,\hat{\lambda}(d\eta)\\
&= \langle \nabla_z \Lambda(\hat{z},\hat{\lambda};\mu_0,\nu_0), z_t - \hat{z}\rangle + o(z_t,\hat{z}) - t\int_a^b \big[\gamma(H(z_t,\eta)) - \sigma([\eta,\infty))\big]\,\hat{\lambda}(d\eta),
\end{aligned}$$

where $o(z_t,\hat{z})/\|z_t - \hat{z}\| \to 0$ as $t \to 0$. By the optimality condition and by the assumption that $\|z_t - \hat{z}\| \leq Lt$, we conclude that

$$\liminf_{t\to 0} \frac{1}{t}\big[\varphi(\mu_t,\nu_t) - \varphi(\mu_0,\nu_0)\big] \geq -\int_a^b \big[\gamma(H(\hat{z},\eta)) - \sigma([\eta,\infty))\big]\,\hat{\lambda}(d\eta). \tag{3.19}$$

Now we use the correspondence between a nonnegative measure $\hat{\lambda}$ on $[a,b]$ and a function $\hat{u} \in \mathscr{U}([a,b])$ defined as follows:

$$\hat{u}(\eta) = \hat{\lambda}([a,\eta]), \quad \eta \in [a,b].$$

Changing the order of integration, as in (3.15), we obtain

$$\int_a^b \gamma(H(\hat{z},\eta))\,\hat{\lambda}(d\eta) = \int \hat{u}(g(\hat{z},v))\,\gamma(dv),$$

$$\int_a^b \sigma([\eta,\infty))\,\hat{\lambda}(d\eta) = \int \hat{u}(y))\,\sigma(y).$$

Using the last two equations, we can rewrite (3.19) as follows:

$$\liminf_{t\to 0} \frac{1}{t}\big[\varphi(\mu_t,\nu_t) - \varphi(\mu_0,\nu_0)\big] \geq -\int \hat{u}(g(\hat{z},v))\,\gamma(dv) + \int \hat{u}(y))\,\sigma(y).$$

As $\hat{\lambda}$ was an arbitrary optimal multiplier, we can take the supremum of the right-hand side over $\hat{u} \in \hat{U}$ to obtain (3.18). □

We point out that the assumption of Lipschitz stability of optimal solutions, $\|z_t - \hat{z}\| \leq Lt$, has an implicit character. In general stability studies, its fulfillment involves appropriate second order sufficient optimality conditions. In our case, due to the nature of the probability distribution functions, such an analysis is very difficult.

Finally, we obtain the directional differentiability result.

COROLLARY 3.7. *Under the assumptions of Theorems* 3.3 *and* 3.6 *the optimal value function is directionally differentiable in the direction* (γ, σ) *with the derivative*

$$\varphi'((\mu_0, \nu_0); (\gamma, \sigma)) = \sup_{\hat{u} \in \hat{U}} \Big\{ -\int \hat{u}\big(g(\hat{z}, v)\big)\, \gamma(dv) + \int \hat{u}(y)\, \sigma(dy) \Big\}.$$

The assumptions simplify considerably if we allow perturbations of the benchmark distribution only.

Acknowledgment. The authors are very grateful to Diethard Klatte and Alexander Shapiro for their insightful remarks and suggestions.

REFERENCES

[1] J. F. BONNANS AND A. SHAPIRO, *Perturbation Analysis of Optimization Problems*, Springer-Verlag, New York, 2000.
[2] R. COMINETTI, *Metric regularity, tangent sets, and second-order optimality conditions*, Appl. Math. Optim., 21 (1990), pp. 265–287.
[3] D. DENTCHEVA AND A. RUSZCZYŃSKI, *Optimization with stochastic dominance constraints*, SIAM J. Optim., 14 (2003), pp. 548–566.
[4] D. DENTCHEVA AND A. RUSZCZYŃSKI, *Optimality and duality theory for stochastic optimization problems with nonlinear dominance constraints*, Math. Program., 99 (2004), pp. 329–350.
[5] D. DENTCHEVA AND A. RUSZCZYŃSKI, *Semi-infinite probabilistic optimization: First order stochastic dominance constraints*, Optimization, 53 (2004), pp. 583–601.
[6] P. C. FISHBURN, *Utility Theory for Decision Making*, Wiley, New York, 1970.
[7] J. HADAR AND W. RUSSELL, *Rules for ordering uncertain prospects*, Amer. Econom. Rev., 59 (1969), pp. 25–34.
[8] R. HENRION AND W. RÖMISCH, *Metric regularity and quantitative stability in stochastic programs with probabilistic constraints*, Math. Program., 84 (1999), pp. 55–88.
[9] A. KIBZUN AND S. URYASEV, *Differentiability of probability function*, Stochastic Anal. Appl., 16 (1998), pp. 1101–1128.
[10] D. KLATTE, *A note on quantitative stability results in nonlinear optimization*, in Proc. 19. Jahrestagung Mathematische Optimierung, K. Lommatzsch, ed., Seminarbericht 90, Humboldt-Universität, Berlin, 1987, pp. 77–86.
[11] D. KLATTE AND R. HENRION, *Regularity and stability in nonlinear semi-infinite optimization*, Nonconvex Optim. Appl., 25 (1998), pp. 69–102.
[12] D. KLATTE AND B. KUMMER, *Nonsmooth Equations in Optimization*, Kluwer Academic, Dordrecht, The Netherlands, 2002.
[13] E. LEHMANN, *Ordered families of distributions*, Ann. Math. Statist., 26 (1955), pp. 399–419.
[14] H. B. MANN AND D. R. WHITNEY, *On a test of whether one of two random variables is stochastically larger than the other*, Ann. Math. Statist., 18 (1947), pp. 50–60.
[15] K. MOSLER AND M. SCARSINI, EDS., *Stochastic Orders and Decision under Risk*, Institute of Mathematical Statistics, Hayward, CA, 1991.
[16] A. PRÉKOPA, *Stochastic Programming*, Kluwer Academic, Dordrecht, Boston, 1995.
[17] J. P. QUIRK AND R. SAPOSNIK, *Admissibility and measurable utility functions*, Rev. Econom. Stud., 29 (1962), pp. 140–146.
[18] S. M. ROBINSON, *Stability theory for systems of inequalities, Part* II: *Differentiable nonlinear systems*, SIAM J. Numer. Anal., 13 (1976), pp. 497–513.
[19] S. M. ROBINSON, *Local epi-continuity and local optimization*, Math. Program., 37 (1987), pp. 208–222.
[20] R. T. ROCKAFELLAR AND R. J.-B. WETS, *Variational Analysis*, Springer-Verlag, Berlin, 1998.

[21] W. Römisch, *Stability of stochastic programming problems*, in Stochastic Programming, A. Ruszczyński and A. Shapiro, eds., Elsevier, Amsterdam, 2003, pp. 483–554.
[22] W. Römisch and R. Schultz, *Stability analysis for stochastic programs*, Ann. Oper. Res., 30 (1991), pp. 241–266.
[23] W. Römisch and R. Schultz, *Distribution sensitivity for certain classes of chance-constrained models with application to power dispatch*, J. Optim. Theory Appl., 71 (1991), pp. 569–588.
[24] A. Ruszczyński and A. Shapiro, eds., *Stochastic Programming*, Elsevier, Amsterdam, 2003.
[25] M. Shaked and J. G. Shanthikumar, *Stochastic Orders and Their Applications*, Academic Press, Boston, MA, 1994.

SIAM J. OPTIM.
Vol. 18, No. 1, pp. 338–363

EFFICIENT LINE SEARCH METHODS FOR CONVEX FUNCTIONS*

EDGAR DEN BOEF† AND DICK DEN HERTOG‡

Abstract. In this paper we propose two new line search methods for convex functions. These new methods exploit the convexity property of the function, contrary to existing methods. The first method is an improved version of the golden section method. For the second method it is proven that after two evaluations the objective gap is at least halved. The practical efficiency of the methods is shown by applying our methods to a real-life bus and buffer size optimization problem and to several classes of convex functions.

Key words. convex optimization, golden section, line search

AMS subject classifications. 65K05, 90C25, 90C56

DOI. 10.1137/04061115X

1. Introduction.

Line searching is an important step for many optimization methods. In practice both exact and approximate line search methods are used. Well-known line search methods are quadratic and cubic interpolation, the golden section method, and backtracking, often combined with clever stopping criteria. For a general overview on line searches we refer to the books by Gill, Murray, and Wright [11], Hiriart-Urruty and Lemarechal [13], and Bazaraa, Sherali, and Shetty [3]. Line search is also an important issue in interior point methods for both linear programming [19] and convex programming [5], [9]. For a good survey on line search techniques within trust-region methods, see Conn, Gould, and Toint [6]. Recent papers on line search techniques in general or on a specific optimization method, like [1], [2], [4], [7], [14], [15], [16], [17], [18], [20], and [21], show that developing and analyzing efficient line search techniques is still an important issue.

The aim in such line search methods is to find a good or a (near) optimal solution w.r.t. an objective or merit function, along a given direction using a minimal number of function evaluations. Especially in the case of black-box functions, where often time-consuming simulation runs, i.e., function evaluations, have to be done, it is desirable to perform as few function evaluations as possible.

Now suppose the (black-box) function is known to be convex (or concave). Note that in practice it happens often that this information is available. For an example see the in-home network problem in section 4. If the function is convex, it has exactly one optimum[1] on a closed domain. This fact is used by the above mentioned methods. However, convexity of a function gives more information. For example, if a function is convex, then using the performed function evaluations, an upper and lower bound can be constructed for the function values. This information can be used to obtain better information on the location of the optimum.

*Received by the editors July 6, 2004; accepted for publication (in revised form) November 5, 2006; published electronically April 17, 2007.

http://www.siam.org/journals/siopt/18-1/61115.html

†Philips Research Laboratories, Eindhoven, The Netherlands. Current address: Quintiq, P.O. Box 264, 5201 AG's-Hertogenbosch, The Netherlands (edgar.den.boef@quintiq.com).

‡Department of Econometrics and OR, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands (D.denHertog@uvt.nl).

[1]In this paper we use the term "optimum" to refer to both the optimal value and the optimal solution; the exact meaning should be clear from the context.

In this paper we will show how this convexity information can be used. We will show that existing methods propose new candidates which may a priori be detected as not optimal or even as nonimproving (w.r.t. the current best point) by using the information of the previous iterations and the convexity property. This means that methods for convex problems that use line searches can be easily improved by adding this simple check for nonimprovement, thereby avoiding unnecessary evaluations.

We will also describe two new methods: the *improved golden section method* and the *triangle section method*. We will show that the improved golden section method is theoretically at least as good as the regular golden section method. For the triangle section method we will show that the objective gap or range of uncertainty, i.e., the difference between the current best known objective value and the lower bound for the optimal value, is at least halved after two function evaluations. It is significant that this result is related to the objective of the optimization, namely, the function values, in contrast to the convergence result of the golden section method, which is related to the function domain values. We also describe the application of our methods to a real-life bus and buffer size optimization problem and to several classes of convex functions. Note that in practice it is much better to have a guarantee for the reduction factor for the function value gap than for the domain gap. We compare the performance of our methods with the performance of the golden section method.

During the revision of this paper we discovered a recent paper written by Gúerin, Marcotte, and Savard [12], in which they describe a method to approximate a univariate convex function. Several elements in their analysis resemble our analysis, although they use derivative information. It is straightforward to extend all our methods for the case in which you also get derivative (or subdifferential) information. In fact then the improvement w.r.t. the golden section is even better. Note, however, that there are many cases in which you do not get derivative information; see, e.g., our practical example in section 4.

In section 2 we show how convexity of a function can be used to reduce the interval in which an optimal solution can be found. For a concave function similar methods can be used. We continue with deriving performance guarantees on the interval reduction in section 3. In section 4 we describe an application involving bus and buffer size optimization in which a nondifferentiable function with computationally expensive function evaluations has to be optimized. We describe experimental results for this application and for other classes of convex functions in section 5. Finally, in section 6 we give our conclusions.

2. Interval reduction using convexity. In this section we describe how to use the convexity of a function to obtain upper and lower bounds for the function values. We show how they can be used to reduce the interval in the function domain in which the minimum can be found, which is called the *interval of uncertainty*. For concave functions a similar method can be used.

Let $f(x)$ be a continuous, univariate, convex function on a closed domain D. Assume that for a given set of points $\mathcal{X} = \{x_1, \ldots, x_n\}$ in D the function values of f are known.

Figure 1(a) gives an example of a convex function f of which six function evaluations are known. As f is convex, $\alpha f(x_i) + (1-\alpha) f(x_j) \geq f(\alpha x_i + (1-\alpha) x_j)$ with $\alpha \in [0,1]$ for each $x_i, x_j \in \mathcal{X}$. Using this property we obtain the piecewise-linear upper bound f^{u} of f with $f^{\mathrm{u}}(x) \geq f(x)$ for all $x \in D$ and $f^{\mathrm{u}}(x) = f(x)$ if $x \in \mathcal{X}$; cf. Figure 1(b). Now we consider the line segments BC and DE and extend them until they intersect at point K, as shown in Figure 1(c). Then, the lines CK and

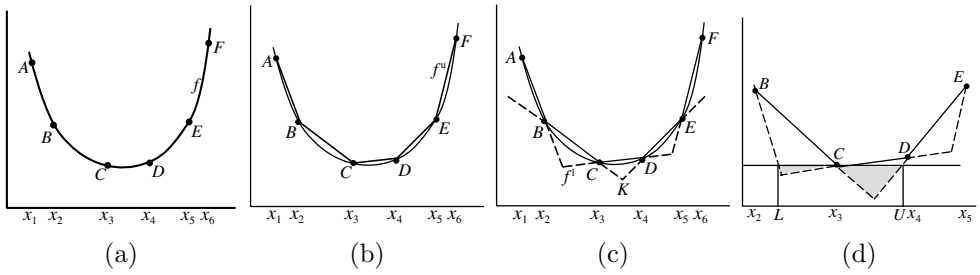


FIG. 1. (a) *Example of a convex function f with six function evaluations.* (b) *A piecewise-linear upper bound based on the convexity property.* (c) *A piecewise-linear lower bound based on the convexity property.* (d) *(Magnified in comparison with* (a)–(c)*.) The optimum lies somewhere in the gray areas, the areas of uncertainty. The interval of uncertainty based on the convexity property is given by $[L, U]$.*

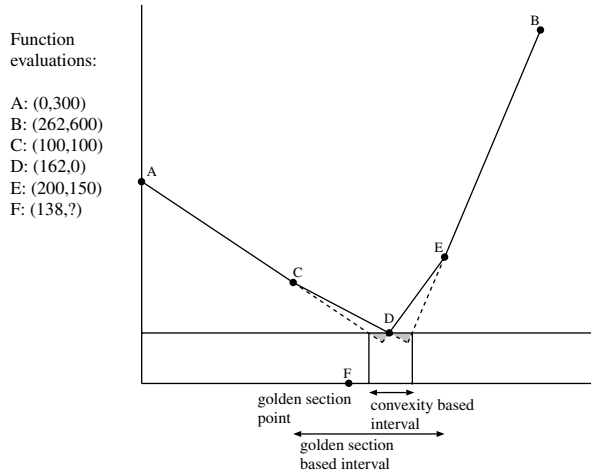


FIG. 2. *Example where the golden section method chooses a point outside the interval of uncertainty for function evaluation. Five function evaluations have already been made following the sequence $A, \ldots, E$. The sixth point for evaluation proposed by golden section is point F. However, the interval of uncertainty comprising the two small gray triangles is located to the right of F.*

KD give a lower bound for the function f between x_3 and x_4. This can be done for any four consecutive points, resulting in the lower bound f^{l} on f given in Figure 1(c) by the dashed line. Now let $x_k \in \mathcal{X}$ be a point with the lowest determined function evaluation, i.e., $f(x_k) \leq f(x)$ for all $x \in \mathcal{X}$. Then the minimum function value of f must lie between $f(x_k)$ and the minimum of $f^{\mathrm{l}}(x)$. The possible locations of the minimum are given in Figure 1(d) by the gray areas, the *areas of uncertainty*. The interval of uncertainty is $[L, U]$.

This shows how the interval of uncertainty can be decreased using the convexity property. The next step in finding the minimum of f is to choose a point for evaluation. Naturally, this should be a point in the interval of uncertainty. Taking one of the existing methods to choose a point, however, does not always give a point in the interval of uncertainty. Figures 2, 3, and 4 show three examples where the golden section method, unit search, and quadratic interpolation, respectively, would evaluate

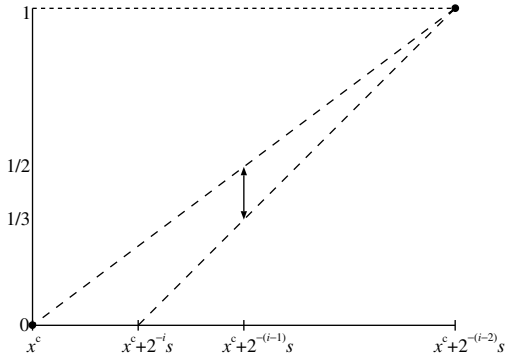


FIG. 3. *Example where unit search chooses a point outside the interval of uncertainty for function evaluation. The known function evaluations are* $f(x^c) = 0$ *and* $f(x^c + 2^{-(i-2)}s) = 1$. *Furthermore, we know* $f(x^c+2^{-(i-1)}s)$. *For the next iteration, unit search would evaluate* $x^c+2^{-i}s$. *However, if* $\frac{1}{3} \leq f(x^c + 2^{-(i-1)}s) \leq \frac{1}{2}$ *holds, then using convexity we know that* $f(x^c + 2^{-i}s) \geq 0$ *and therefore will not lead to a new minimum. Unit search can be improved by choosing each time the point in the middle of the interval of uncertainty based on convexity.*

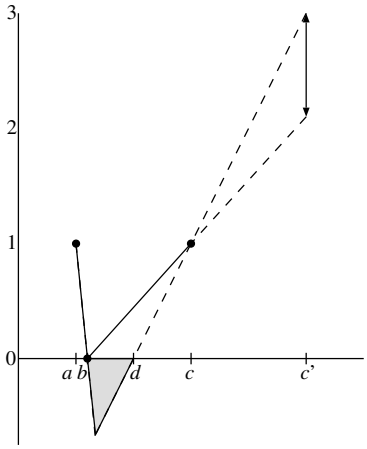


FIG. 4. *Example where quadratic interpolation chooses a point outside the interval of uncertainty for function evaluation. The points with known function evaluations are* $a = 0$, $b = \frac{1}{10}$, *and* $c = 1$, *with* $f(a) = f(c) = 1$ *and* $f(b) = 0$. *Furthermore, we know* $f(c')$. *Quadratic interpolation would take* $d = \frac{1}{2}$ *as the point for a new function evaluation. However, if* $2\frac{1}{9} \leq f(c') \leq 3$ *holds, then using convexity it is clear that* $f(d) \geq 0$ *and therefore will not lead to a new minimum.*

the function at a point that is outside the interval of uncertainty. In section 3 we discuss some strategies for choosing a new point. Finally, the interval of uncertainty may even be further decreased if, for an evaluated point, the gradient of the function is known. However, we leave this for future research.

3. Function evaluation strategies. In this section we discuss two strategies for choosing a new point that use the convexity property. In section 3.1 we show how to choose a new point for evaluation when the focus lies on the reduction of the interval of uncertainty. In section 3.2 we show how to choose a new point such that

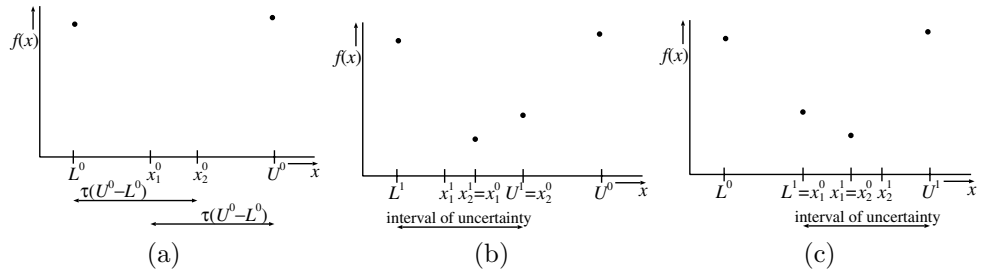


FIG. 5. *Starting the golden section method.* (a) *First two points,* x_1^0 *and* x_2^0*, are chosen in the interior of the starting interval of uncertainty,* $[L^0, U^0]$*. Next, the function is evaluated at these points, and depending on which one is lower, the interval of uncertainty is adjusted; see* (b) *and* (c).

the *range of uncertainty*, i.e., the interval consisting of all possible function values for the minimum, is at least halved for each two new function evaluations. Finally, in section 3.3 we shortly describe how known piecewise linearity of a function can be used to terminate the search procedure.

3.1. Function domain reduction. For the reduction of the interval of uncertainty in the function domain we have taken the golden section method and improved it with the reduction that follows from the convexity property, as described in the previous section. The golden section method chooses new points for function evaluation in such a way that the interval of uncertainty can be decreased by a constant factor $\tau = (\sqrt{5}-1)/2$ in each iteration. Figure 5 shows an example. Let $[L^0, U^0]$ be the initial interval of uncertainty. Then golden section method chooses the following two interior points x_1^0 and x_2^0 for evaluation: $x_1^0 = U^0 - \tau(U^0 - L^0)$ and $x_2^0 = L^0 + \tau(U^0 - L^0)$. Now, suppose x_1^0 has the lowest function evaluation. Then the new interval of uncertainty $[L^1, U^1]$ is equal to $[L^0, x_2^0]$. Furthermore, the new interior points to be evaluated are $x_1^1 = U^1 - \tau(U^1 - L^1)$ and $x_2^1 = L^1 + \tau(U^1 - L^1)$. However, as the interior points are chosen using the golden section factor τ, it follows that $x_2^1 = x_1^0$. Therefore, only x_1^1 has to be evaluated for the next step. Similarly, if x_2^0 has the lowest function evaluation, then $L^1 = x_1^0$ and $x_1^1 = x_2^0$.

The improved golden section method, shown in Algorithm 1, now works as follows. It starts in the same way as the regular golden section method. Let $[L, U]$ be the interval of uncertainty with two interior points x_1 and x_2 such that $x_2 - L = U - x_1 = \tau(U - L)$. We first consider the case in which the minimum function evaluation occurs at one of the interval boundaries, e.g., at L, so $f(L) = \min\{f(L), f(x_1), f(x_2), f(U)\}$. Then the new interval of uncertainty according to the golden section method is given by $[L, x_1]$. However, using the convexity property we can obtain a smaller interval of uncertainty $[L, U']$ with $U' \leq x_1$ for which two new interior points are selected. If $f(U) = \min\{f(L), f(x_1), f(x_2), f(U)\}$, then similarly a new interval $[L', U]$ can be obtained. Note that the minimum function evaluation can occur only at L or U if it has not occurred yet at an internal point of the interval of uncertainty. Furthermore, by evaluating the function f first at the boundaries L and U and then at the interior point closest to the boundary with the lowest function value, e.g., x_1 if $f(L) < f(U)$, the function does not need to be evaluated at the other interior point if the lowest function value is not improved by the new evaluation; see Figure 6 for an example and Algorithm 2 for a description of the algorithm loop during initialization.

Now we assume that x_1 has the lowest function value, i.e., $f(x_1) \leq f(x_2)$; if $f(x_2) < f(x_1)$ holds, then we can follow a strategy analogous to what we describe

Algorithm 1. The improved golden section method.

Input: Interval of uncertainty $[L, U]$, stop criterion ε,
Output: Interval of uncertainty $[L, U]$ with $U - L < \varepsilon$,
Parameter: Golden section $\tau = \frac{\sqrt{5}-1}{2}$.

Initialization:
Perform initialization loop of Algorithm 2 to obtain a new interval $[L, U]$ with interior point x;
Loop:
while $U - L > \varepsilon$ **do**
if $x = U - \tau(U - L)$ **then**
$x_1 := x$, $x_2 := L + \tau(U - L)$;
Determine $f(x_2)$;
else $(x = L + \tau(U - L))$
$x_1 := U - \tau(U - L)$, $x_2 := x$;
Determine $f(x_1)$;
endif
Determine new interval of uncertainty $[L', U']$ using convexity property;
Stretch $[L', U']$ to $[\tilde{L}, \tilde{U}]$ according to (1) and (2) to maintain golden section property;
Take as new interval of uncertainty $[L, U] := [\tilde{L}, \tilde{U}]$;
If $x_1 \in (L, U)$, then $x := x_1$, else if $x_2 \in (L, U)$, then $x := x_2$;
endwhile

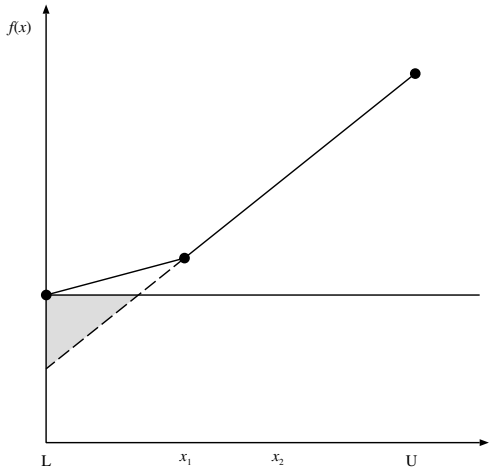


FIG. 6. *Initialization of improved golden section method where the minimum still lies at L after evaluation at x_1. As x_2 lies outside the new interval of uncertainty, it does not need to be evaluated anymore.*

here. The new interval of uncertainty using the golden section method is now given by $[L, x_2]$. Using the convexity property we can obtain a smaller interval of uncertainty $[L', U']$ for which $L' \geq L$ and $U' \leq x_2$ holds. For now we assume that x_1 is an interior point of $[L', U']$; the possibility that x_1 is not an interior point is handled later in this section.

Algorithm 2. Loop at initialization of the improved golden section method to obtain the lowest function evaluation at an interior point of the interval of uncertainty.

Input: Interval of uncertainty $[L, U]$, stop criterion ε,
Output: Interval of uncertainty $[L, U]$ with minimum at interior point x or $U - L < \varepsilon$,
Parameter: Golden section $\tau = \frac{\sqrt{5}-1}{2}$.

Initialization:
Determine $f(L)$ and $f(U)$;
Loop:
repeat
if $f(L) < f(U)$ **then**
$x := U - \tau(U - L)$;
else
$x := L + \tau(U - L)$;
Determine $f(x)$;
Determine new interval of uncertainty $[L, U]$ using convexity property;
until $U - L < \varepsilon$ or $f(x) \leq f(L)$ and $f(x) \leq f(U)$;

The golden section method would choose a new point $x_3 = x_2 - \tau(x_2 - L)$ so that $x_2 - x_3 = x_1 - L$. However, if we replace $[L, x_2]$ by $[L', U']$ and then choose $x_3 = U' - \tau(U' - L')$, $U' - x_3$ is generally not equal to $x_1 - L'$, meaning that the two interior points x_1 and x_3 do not satisfy the golden section property w.r.t. the interval of uncertainty $[L', U']$. Therefore, we will stretch the interval $[L', U']$ to a new interval $[\tilde{L}, \tilde{U}]$ such that the golden section property can be maintained for the new point to be evaluated. We distinguish four possibilities for this:

$$\begin{array}{ll} \text{(a)} & x_1 \leq U' - \tau(U' - L'), \\ \text{(b)} & U' - \tau(U' - L') < x_1 < \frac{1}{2}(U' + L'), \\ \text{(c)} & \frac{1}{2}(U' + L') \leq x_1 < L' + \tau(U' - L'), \\ \text{(d)} & x_1 \geq L' + \tau(U' - L'). \end{array} \tag{1}$$

Figure 7 shows an example of these four possibilities. The corresponding stretched intervals are the following:

$$\begin{array}{lll} \text{(a)} & \tilde{L} = U' - \frac{1}{\tau}(U' - x_1), & \tilde{U} = U', \\ \text{(b)} & \tilde{L} = L', & \tilde{U} = \frac{1}{1-\tau}(x_1 - \tau L'), \\ \text{(c)} & \tilde{L} = \frac{1}{1-\tau}(x_1 - \tau U'), & \tilde{U} = U', \\ \text{(d)} & \tilde{L} = L', & \tilde{U} = L' + \frac{1}{\tau}(x_1 - L'). \end{array} \tag{2}$$

The next lemma states that the obtained stretched interval is not larger than the interval of uncertainty obtained with the regular golden section method.

LEMMA 1. *The stretched interval, $[\tilde{L}, \tilde{U}]$, is not larger than the interval of uncertainty of the regular golden section method, $[L, x_2]$, i.e.,*

$$\tilde{U} - \tilde{L} \leq x_2 - L \, .$$

Proof. For (a),(c), and (d) we prove that $\tilde{U} - \tilde{L} \leq x_2 - L$ holds by showing that $\tilde{L} \geq L$ and $\tilde{U} \leq x_2$. For (b) it is possible that $\tilde{U} > x_2$. In the following derivations we use the fact that $\frac{1-\tau}{\tau} = \tau$.

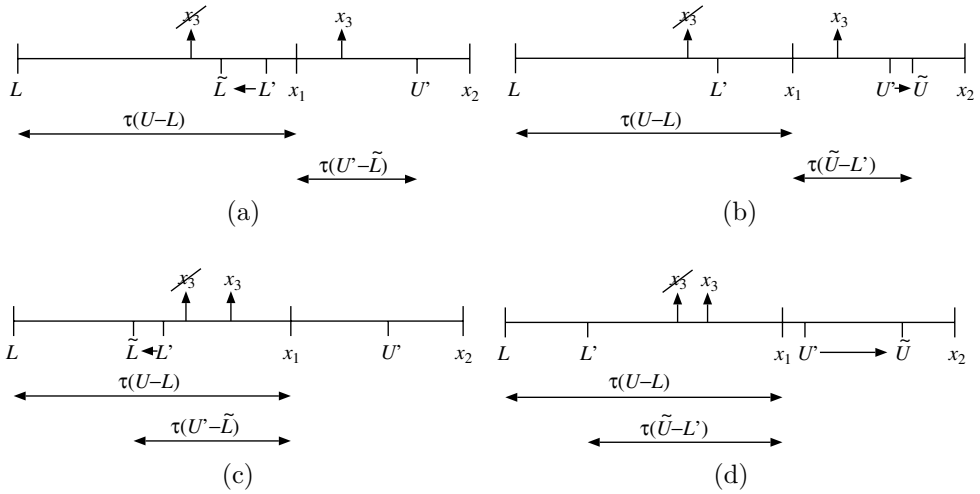


FIG. 7. *Stretching the interval of uncertainty obtained with the convexity property such that the golden section property is maintained for the new function evaluation. The four figures correspond to the four different possibilities.*

(a) As $\tilde{U} = U'$ it follows that $\tilde{U} \leq x_2$. For $\tilde{L}$ we can derive

$$\begin{aligned}
\tilde{L} &= U' - \frac{1}{\tau}(U' - x_1) \\
&= \left(1 - \frac{1}{\tau}\right) U' + \frac{1}{\tau}(L + \tau(x_2 - L)) \\
&= x_2 - \frac{1-\tau}{\tau}(U' - L) \\
&= x_2 - \tau(U' - L) \\
&\geq x_2 - \tau(x_2 - L) \\
&> x_2 - (x_2 - L) \\
&= L.
\end{aligned}$$

(b) From $x_1 \leq \frac{1}{2}(L' + U')$ it follows that $\frac{1}{2}(x_1 - L') \leq \frac{1}{2}(U' - x_1)$, and therefore $x_1 - L' \leq U' - x_1 \leq x_2 - x_1$. We can now make the following derivation:

$$\begin{aligned}
\tilde{U} - \tilde{L} &= \frac{x_1 - \tau L'}{1-\tau} - L' \\
&= \frac{x_1 - L'}{1-\tau} \\
&\leq \frac{x_2 - x_1}{1-\tau} \\
&= \frac{x_2 - (L + \tau(x_2 - L))}{1-\tau} \\
&= x_2 - L.
\end{aligned}$$

(c) As $\tilde{U} = U'$ it follows that $\tilde{U} \leq x_2$. For $\tilde{L}$ we can derive

$$\begin{aligned}\tilde{L} &= \frac{1}{1-\tau}(x_1 - \tau U') \\ &= \frac{1}{1-\tau}(L + \tau(x_2 - L)) - \frac{\tau}{1-\tau}U' \\ &= L + \frac{1}{\tau}(x_2 - U') \\ &\geq L.\end{aligned}$$

(d) As $\tilde{L} = L'$ it follows that $\tilde{L} \geq L$. For $\tilde{U}$ we can derive

$$\begin{aligned}\tilde{U} &= L' + \frac{1}{\tau}(x_1 - L') \\ &= \left(1 - \frac{1}{\tau}\right) L' + \frac{1}{\tau}(L + \tau(x_2 - L)) \\ &= x_2 - \left(\frac{1-\tau}{\tau}\right)(L' - L) \\ &= x_2 - \tau(L' - L) \\ &\leq x_2. \qquad \square\end{aligned}$$

This leads to the following strategy for choosing a new point for evaluation.

Improved golden section strategy, $f(x_1) \leq f(x_2)$. Determine the stretched interval $[\tilde{L}, \tilde{U}]$ as described above. Choose the new point x_3 for function evaluation as follows for the four previously distinguished possibilities:

(a) $x_1 \leq U' - \tau(U' - L')$, $\quad x_3 = \tilde{L} + \tau(\tilde{U} - \tilde{L}) = U' - \tau(U' - x_1)$.
(b) $U' - \tau(U' - L') < x_1 < \frac{1}{2}(L' + U')$, $\quad x_3 = \tilde{L} + \tau(\tilde{U} - \tilde{L}) = \frac{1}{\tau}x_1 - \tau L'$.
(c) $\frac{1}{2}(L' + U') \leq x_1 < L' + \tau(U' - L')$, $\quad x_3 = \tilde{U} - \tau(\tilde{U} - \tilde{L}) = \frac{1}{\tau}x_1 - \tau U'$.
(d) $x_1 \geq L' + \tau(U' - L')$, $\quad x_3 = \tilde{U} - \tau(\tilde{U} - \tilde{L}) = L' + \tau(x_1 - L')$.

In the following theorem we show that the improved golden section strategy performs at least as well as the regular golden section method, while ensuring that the new point chosen for function evaluation lies in the interval of uncertainty.

THEOREM 1. *Let $[L, U]$ be the interval of uncertainty, and let $x_1, x_2 \in [L, U]$ such that $U - x_1 = x_2 - L = \tau(U - L)$, i.e., x_1 and x_2 satisfy the golden section property. Without loss of generality (w.l.o.g.) assume that $f(x_1) \leq f(x_2)$. Then the following hold:*

(a) *The stretched interval of uncertainty $[\tilde{L}, \tilde{U}]$ obtained by the improved golden section method is at least a factor τ smaller than the starting interval of uncertainty $[L, U]$, i.e.,*

$$\tilde{U} - \tilde{L} \leq \tau(U - L)\ .$$

(b) *Let x_3 be the new point for function evaluation chosen according to the improved golden section strategy. Then, x_3 lies in the interval of uncertainty, i.e.,*

$$x_3 \in [L', U']\ .$$

Proof. (a) Using the golden section method, the interval of uncertainty $[L, U]$ with function evaluations at $x_1 = U - \tau(U - L)$ and $x_2 = L + \tau(U - L)$ reduces to

interval $[L, x_2]$ of size $\tau(U - L)$. The new point for function evaluation x_3 is then chosen such that the golden section property is maintained, i.e., $x_3 = x_2 - \tau(x_2 - L)$.

Using the improved golden section strategy, the new point for function evaluation is chosen to be either $x_3 = \tilde{U} - \tau(\tilde{U} - \tilde{L})$ or $x_3 = \tilde{L} + \tau(\tilde{U} - \tilde{L})$. Expressing the other internal point x_1 in $\tilde{L}$ and $\tilde{U}$ by rewriting the expressions for $\tilde{L}$ and $\tilde{U}$ given in (1) and (2) gives $x_1 = \tilde{L} + \tau(\tilde{U} - \tilde{L})$ for (c) and (d) and $x_1 = \tilde{U} - \tau(\tilde{U} - \tilde{L})$ for (a) and (b). So for the improved golden section strategy, the golden section property is maintained for the interval $[\tilde{L}, \tilde{U}]$. Lemma 1 states that $\tilde{U} - \tilde{L} \leq x_2 - L$ and thus $\tilde{U} - \tilde{L} \leq \tau(U - L)$. Therefore, the starting interval of uncertainty is reduced by at least a factor of τ. Furthermore, as the new interval of uncertainty has the golden section property, the same reduction factor is guaranteed for the following function evaluations.

(b) Now we show that $x_3 \in [L', U']$. For (a), $x_3 = U' - \tau(U' - x_1)$, so $x_3 < U'$ is obvious. Furthermore, using $\tau < 1$ and $x_1 > L'$ it follows that $x_3 > L'$. Likewise, for (d) $x_3 \in [L', U']$ holds. For (b) we can make the following derivations:

$$\begin{aligned} x_3 &= \frac{1}{\tau}x_1 - \tau L' \\ &\leq \frac{L' + U'}{2\tau} - \tau L' \\ &= \frac{U'}{\tau} - \frac{(U' - L')}{2\tau} - \tau U' + \tau(U' - L') \\ &= U' - \frac{1 - 2\tau^2}{2\tau}(U' - L') \\ &< U', \end{aligned}$$

$$\begin{aligned} x_3 &= \frac{1}{\tau}x_1 - \tau L' \\ &> \frac{1}{\tau}(U' - \tau(U' - L')) - \tau L' \\ &= \frac{1 - \tau}{\tau}U' + (1 - \tau)L' \\ &= \tau U' + (1 - \tau)L' \\ &> L'. \end{aligned}$$

In a symmetrical manner it can be shown for (c) that $x_3 \in [L', U']$. ∎

For completeness, we give the formula for the new point x_3 for evaluation in case $f(x_2) < f(x_1)$.

Improved golden section strategy, $f(x_2) < f(x_1)$. Determine the stretched interval $[\tilde{L}, \tilde{U}]$. Choose the new point x_3 for function evaluation as follows:

(a) $x_2 \geq L' + \tau(U' - L')$, $x_3 = L' + \tau(x_2 - L')$.
(b) $\frac{1}{2}(L' + U') \leq x_2 < L' + \tau(U' - L')$, $x_3 = \frac{1}{\tau}x_2 - \tau U'$.
(c) $U' - \tau(U' - L') < x_2 < \frac{1}{2}(L' + U')$, $x_3 = \frac{1}{\tau}x_2 - \tau L'$.
(d) $x_2 \leq U' - \tau(U' - L')$, $x_3 = U' - \tau(U' - x_2)$.

In a way analogous to the above, we can show that this reduces the interval of uncertainty by at least a factor of τ.

Finally, we consider the possibility in which $[L', U']$ does not have one of the previous interior points x_1, x_2 as an interior point. This can happen when $f(x_1) =$

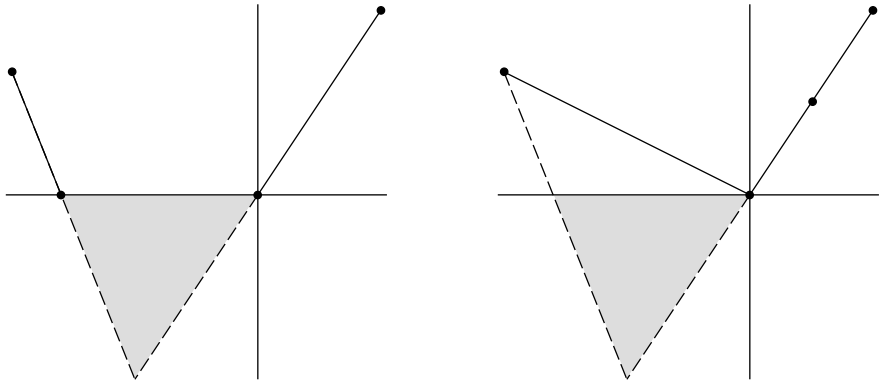


FIG. 8. *Two examples when the improved golden section method leads to an interval of uncertainty without an interior point. In* (a) *the two interior points of the previous interval of uncertainty both have the same function value. In* (b) *the lowest function evaluation lies on one line with two other function evaluations.*

$f(x_2)$ or when at least three consecutive points that have already been evaluated lie on one line (see Figure 8); in the latter case the function f is at least partially piecewise linear. Then the interval $[L', U']$ lies between x_1 and x_2, i.e., $[L', U'] \subseteq [x_1, x_2]$ with either $L' = x_1$ or $U' = x_2$ or both. Two new function evaluations are now required to decrease the interval further by at least a factor of τ. However, as $x_2 - x_1 = \tau^2(U - L)$, the previous function evaluation results in a reduction of τ^2. Therefore, the combined reduction of the previous function evaluation and the new function evaluation will also be at least τ^2, giving an average reduction of at least a factor of τ for each function evaluation.

We now choose one new point using the golden section property, i.e., either $x_3 = x_2 - \tau(x_2 - x_1)$ or $x_3 = x_1 + \tau(x_2 - x_1)$, which ensures that a reduction of at least τ can be guaranteed for the following function evaluations. The resulting function evaluation of x_3 is used to update the interval of uncertainty, and we choose another point according to the described improved golden section strategy.

Now we have shown how the convexity property can be used to choose new points for evaluation such that the interval of uncertainty is reduced by at least the same factor as for the golden section method. However, in practice the reduction will be larger as we show with empirical results in section 5. In this section we continue with a method that decreases the range of uncertainty by at least a factor of 1/2 after two new function evaluations.

3.2. Function range reduction. As is shown in Figure 1(d) the convexity property can be used to obtain upper and lower bounds for the function value of each point in the interval of uncertainty. As these upper and lower bounds tighten for each new function evaluation, they can be used for a strategy that guarantees a reduction of the range of uncertainty instead of the interval of uncertainty. Figure 9 depicts the area in which the optimum can be found, together with the points corresponding to the function evaluations and the interval of uncertainty.

Let M be the point with the lowest function evaluation so far, $f(M)$. Then $L' \leq M \leq U'$. Now we define Δf_1^k as the height of the triangle in the area of uncertainty between L' and M after k function evaluations, and we define Δf_2^k as the height of the triangle between M and U'. We can express Δf_1^k and Δf_2^k using known

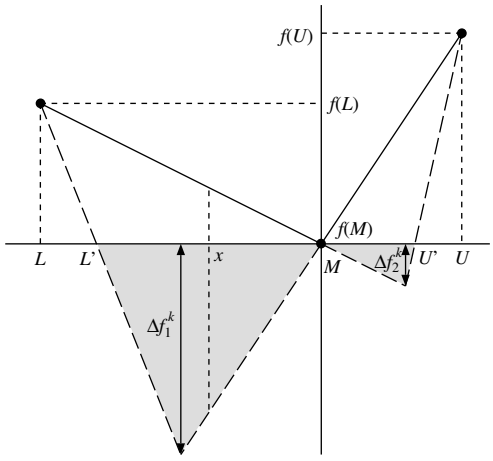


FIG. 9. *The areas of uncertainty. The three points with their function evaluations are given by L, M, and U. The interval of uncertainty in the function domain begins at L' and ends at U'. The height of the two areas after k function evaluations is given by Δf_1^k and Δf_2^k. The new point for function evaluation is x.*

function values and the interval of uncertainty, as derived in Appendix A, which gives the following formulas:

$$\Delta f_1^k = \frac{(M - L')(f(L) - f(M))(f(U) - f(M))}{(f(L) - f(M))(U - M) + (f(U) - f(M))(L' - L)}, \tag{3}$$

$$\Delta f_2^k = \frac{(U' - M)(f(L) - f(M))(f(U) - f(M))}{(f(L) - f(M))(U - U') + (f(U) - f(M))(M - L)}. \tag{4}$$

The range of uncertainty is now given by the maximum height of the area of uncertainty, i.e., $\max\{\Delta f_1^k, \Delta f_2^k\}$. The point x we choose for function evaluation now lies in the middle of the area with the largest height, i.e.,

$$x = \begin{cases} \frac{1}{2}(L' + M) & \text{if} \quad \Delta f_1^k \geq \Delta f_2^k, \\ \frac{1}{2}(M + U') & \text{if} \quad \Delta f_1^k < \Delta f_2^k. \end{cases} \tag{5}$$

We refer to the method that chooses a new point for function evaluation according to (5) as the *triangle section method*; see also Algorithm 3 for an overview of the method. A greedy strategy is to take the point where the lower bound is minimal. However, it can be easily shown that the performance bounds for this greedy strategy are worse than the performance bounds for the triangle section strategy which we give in this paper.

In the remainder of this section we normalize, w.l.o.g., the function values, the size of the interval of uncertainty, and the range of uncertainty in the following way:

$$M = 0, \qquad f(M) = 0, \qquad \Delta f_1^k = 1, \qquad L' = -1.$$

After substituting these values into the expression for Δf_1^k, it follows that $f(U) = \frac{f(L)U}{1+L+f(L)}$ if $1 + L + f(L) \neq 0$. The values of L, $f(L)$, U, and U' then determine the exact situation. For the ease of notation we make the following substitutions, as

Algorithm 3. The triangle section method.

Input: Interval of uncertainty $[L, U]$, stop criterion ε,
Output: Interval of uncertainty $[L, U]$ with size of range of uncertainty $\max\{\Delta f_1^k, \Delta f_2^k\} < \varepsilon$;

Initialization: Take $x = (U - L)/2$ and $k := 0$,
Loop:
repeat
$k := k + 1$;
Determine $f(x)$;
Determine new interval of uncertainty $[L', U']$ using convexity property;
Determine $\Delta f_1^k, \Delta f_2^k$;
Take as new interval of uncertainty $[L, U] := [L', U']$;
Determine new point x according to (5)
until $\max\{\Delta f_1^k, \Delta f_2^k\} < \varepsilon$

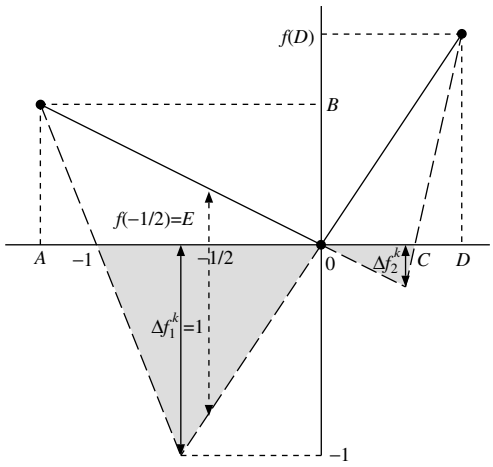


FIG. 10. *The areas of uncertainty with normalization of the function and interval values. The point* 0 *has the lowest function evaluation of* 0. *The size of the range of uncertainty is given by* Δf_1^k, *which is set to* 1. *The lower bound of the interval of uncertainty in the function domain is equal to* -1; *thus, the new point for function evaluation is* $-\frac{1}{2}$.

shown in Figure 10:

$$\begin{aligned} A &= L, \\ B &= f(L), \\ C &= U', \\ D &= U, \\ E &= f(x). \end{aligned}$$

Furthermore, w.l.o.g. we assume that $\Delta f_1^k \geq \Delta f_2^k$. The new point for evaluation then is $x = -\frac{1}{2}$.

For the values of A, B, C, D, and E we can derive the following properties:

(i) The lower corner of the left area should be to the left of M, i.e., $-D/f(D) \leq 0$. As $f(D) = f(U) = f(L)U/[1+L+f(L)] = BD/[1+A+B]$, if $1+A+B \neq 0$, we have $-[1+A+B]/B \leq 0$ and $D \neq 0$. Since $B = f(L) > 0$, it follows that $1 + A + B \geq 0$.

(ii) $L \leq L'$, i.e., $A \leq -1$ or $-1 - A \geq 0$ or $1 + A \leq 0$.

(iii) $U \geq U' \geq 0$, i.e., $D \geq C \geq 0$.

(iv) $\Delta f_2^k \leq \Delta f_1^k = 1$. Substitution of Δf_2^k gives $\Delta f_2^k = \frac{BCD}{(D-C)(1+A+B)-AD}$. So $BCD \leq (D - C)(1 + A + B) - AD$.

(v) Let $f^{\mathrm{u}}(x)$ denote the upper bound for point x, i.e., the line $(A, B) - (0, 0)$. Then $f(x) \leq f^{\mathrm{u}}(x)$ should hold, i.e., $E \leq -\frac{B}{2A}$ or $-2AE \leq B$.

(vi) Let $f^{\mathrm{l}}(x)$ denote the lower bound for point x, i.e., the line parts below the $y = 0$ line of the lines (A, B)–$(-1, 0)$ and $(0, 0)$–$(D, f(D))$. Then $f(x) \geq f^{\mathrm{l}}(x)$ should hold. The value of $f^{\mathrm{l}}(x)$ depends on whether the lowest corner of the left area lies to the left or to the right of x, i.e., $f^{\mathrm{l}}(x) = \max\{-\frac{B}{2(1+A+B)}, -\frac{B}{2(-1-A)}\}$. This gives $B \geq -2E(1 + A + B)$ and $B \geq -2E(-1 - A)$. Notice that this also holds if $1 + A + B = 0$ or $-1 - A = 0$.

We now show in Lemma 2 that in the special case the area of uncertainty consists of one triangle, i.e., $\Delta f_1^k = 0$ or $\Delta f_2^k = 0$; the triangle section method at least halves the range of uncertainty for a new function evaluation.

LEMMA 2. *Let a new function evaluation x_{k+1} be chosen according to* (5)*, and let either $\Delta f_1^k = 0$ or $\Delta f_2^k = 0$. Then the range of uncertainty decreases by at least a factor of $\frac{1}{2}$ after the function evaluation $f(x_{k+1})$ is known, i.e.,*

$$\max\{\Delta f_1^{k+1}, \Delta f_2^{k+1}\} \leq \frac{1}{2} \max\{\Delta f_1^k, \Delta f_2^k\} .$$

Proof. W.l.o.g. we assume that $\Delta f_2^k = 0$. We distinguish four possibilities for $f(x_{k+1})$; cf. Figure 11(a).

1. $f(x_{k+1}) = f^{\mathrm{u}}(x_{k+1})$; see Figure 11(b). Then the upper bound is equal to the lower bound for all points in $[L, U]$, and $M = 0$ is the optimum with function value $f(M) = 0$. Thus $\Delta f_1^{k+1} = 0$.
2. $f^{\mathrm{u}}(x_{k+1}) > f(x_{k+1}) \geq 0$; see Figure 11(c). If we write Δf_1^{k+1} as an expression of A, B, C, D, and E, we get

$$\Delta f_1^{k+1} = \frac{B(\frac{1}{2}B + AE)}{B(-\frac{1}{2} - A) + (B - E)(1 + A + B)}. \tag{6}$$

   Now we need to show that $\Delta f_1^{k+1} \leq \frac{1}{2}$ or $\frac{1}{2} - \Delta f_1^{k+1} \geq 0$. From properties (i), (ii), and (v) it follows that $B(-\frac{1}{2} - A) + (B - E)(1 + A + B) > 0$. So we need to show that $\frac{1}{2}B(-\frac{1}{2} - A) + \frac{1}{2}(B - E)(1 + A + B) - B(\frac{1}{2}B + AE) \geq 0$. Rewriting the left part of this inequality gives $\frac{1}{4}B + \frac{1}{2}E((-1 - A)(1 + B) - AB)$. Since $B, E, (-1 - A) \geq 0$ the inequality holds, and $\Delta f_1^{k+1} \leq \frac{1}{2}$.
3. $0 > f(x_{k+1}) > f^{\mathrm{l}}(x_{k+1})$; see Figure 11(d). Then both $\Delta f_1^{k+1} > 0$ and $\Delta f_2^{k+1} > 0$ will hold. If we write Δf_1^{k+1} as an expression of A, B, C, D, and E, we get

$$\Delta f_1^{k+1} = \frac{-E(\frac{1}{2}B + E(-1 - A))}{\frac{1}{2}B - E(-1 - A)}. \tag{7}$$

   Now we show that $\frac{1}{2} - \Delta f_1^{k+1} \geq 0$. As $B > 0$, $E < 0$, and $-1 - A > 0$, this can be done by showing that $\frac{1}{2}(\frac{1}{2}B - E(-1 - A)) - (-E(\frac{1}{2}B + E(-1 - A))) \geq 0$.

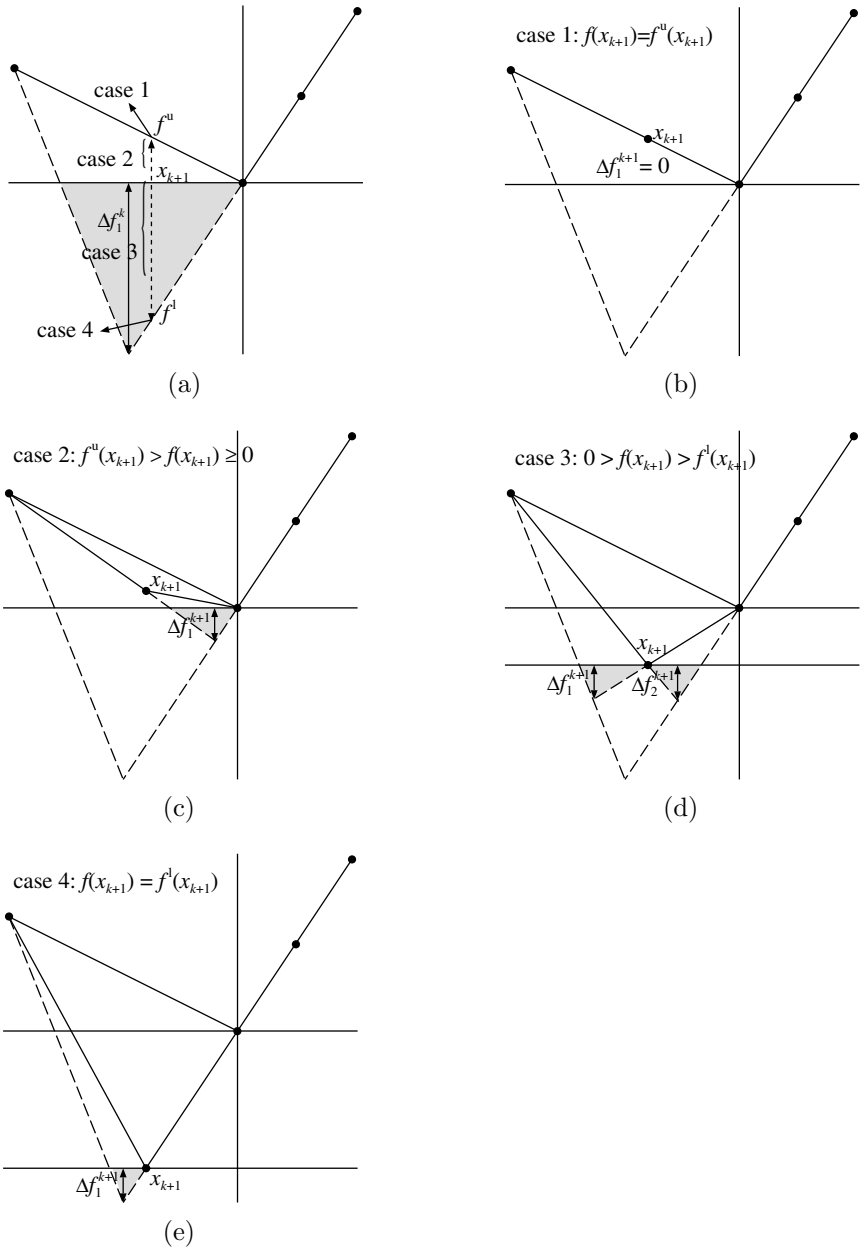


FIG. 11. *Decrease of range of uncertainty when* $\Delta f_2^k = 0$. (a) *Four possibilities are distinguished for the new function evaluation* $f(x_{k+1})$. (b) $f(x_{k+1}) = f^{\mathrm{u}}(x_{k+1})$. (c) $f^{\mathrm{u}}(x_{k+1}) > f(x_{k+1}) \geq 0$. (d) $0 > f(x_{k+1}) > f^{\mathrm{l}}(x_{k+1})$. (e) $f(x_{k+1}) = f^{\mathrm{l}}(x_{k+1})$.

Rewriting the left part of this inequality and using $B \geq -2E(1+A+B)$ from property (vi) gives $(-1-A)E^2 + \frac{1}{2}E(1+A+B) + \frac{1}{4}B \geq (-1-A)E^2 \geq 0$. If we write Δf_2^{k+1} as an expression of A, B, C, D, and E, we get

$$\Delta f_2^{k+1} = \frac{E(B-E)(1+A+B) + \frac{1}{2}B(B-E)}{(B-E)(1+A+B) + B(-\frac{1}{2}-A)}. \tag{8}$$

As $B > E$, $1+A+B > 0$, $B > 0$, and $-\frac{1}{2} - A > -1 - A \geq 0$, it suffices to show that $\frac{1}{2}((B-E)(1+A+B)+B(-\frac{1}{2}-A))-(E(B-E)(1+A+B)+\frac{1}{2}B(B-E)) \geq 0$. Rewriting the left part of this inequality and using $B \geq -2E(-1-A)$ from property (vi) gives $-\frac{1}{2}E(1+A)+\frac{1}{4}B-E(B-E)(1+A+B) \geq -\frac{1}{2}E(1+A)+\frac{1}{4}(-2E(-1-A))-E(B-E)(1+A+B) = -E(B-E)(1+A+B) \geq 0$.

4. $f(x_{k+1}) = f^{\mathrm{l}}(x_{k+1})$; see Figure 11(e). Then $\Delta f_2^{k+1} = 0$, and $\Delta f_1^{k+1} = \Delta f_1^k - (f(M) - f^{\mathrm{l}}(x^{k+1})) = \Delta f_1^k + f^{\mathrm{l}}(x_{k+1})$. So, for $\Delta f_1^{k+1} \leq \frac{1}{2}\Delta f_1^k = \frac{1}{2}$ we need to show that $-f^{\mathrm{l}}(x_{k+1}) - \frac{1}{2} \geq 0$. The lower bound is given by $f^{\mathrm{l}}(x_{k+1}) = \max\{-\frac{B}{2(1+A+B)}, -\frac{B}{2(-1-A)}\}$. If $f^{\mathrm{l}}(x_{k+1}) = -\frac{B}{2(1+A+B)}$ we have

$$\frac{B}{2(1+A+B)} - \frac{1}{2} = \frac{B-(1+A+B)}{2(1+A+B)} = \frac{-1-A}{2(1+A+B)} \geq 0.$$

If $f^{\mathrm{l}}(x_{k+1}) = -\frac{B}{2(-1-A)}$ we have

$$\frac{B}{2(-1-A)} - \frac{1}{2} = \frac{B-(-1-A)}{2(-1-A)} = \frac{1+A+B}{2(-1-A)} \geq 0. \qquad \square$$

We use Lemma 2 to show that the triangle section method at least halves the range of uncertainty after two new function evaluations.

THEOREM 2. *Let each new function evaluation x_{k+1} be chosen according to* (5). *Then the range of uncertainty is at least halved after two function evaluations, i.e., for all k,*

$$\max\{\Delta f_1^{k+2}, \Delta f_2^{k+2}\} \leq \frac{1}{2}\max\{\Delta f_1^k, \Delta f_2^k\}\,.$$

Proof. W.l.o.g. we assume that $\Delta f_1^k \geq \Delta f_2^k$. We distinguish three possibilities for the function value $f(x_{k+1})$.

1. $f(x_{k+1}) = f^{\mathrm{u}}(x_{k+1})$; see Figure 12(b). Then for all $y \in [L, M]$ we have $f^{\mathrm{u}}(y) = f^{\mathrm{l}}(y)$ and $\Delta f_1^{k+1} = 0$. Furthermore, $\Delta f_2^{k+1} = \Delta f_2^k$. Now it follows from Lemma 2 that $\max\{\Delta f_1^{k+2}, \Delta f_2^{k+2}\} \leq \frac{1}{2}\Delta f_2^{k+1} \leq \frac{1}{2}\max\{\Delta f_1^k, \Delta f_2^k\}$.
2. $0 \geq f(x_{k+1}) \geq f^{\mathrm{l}}(x_{k+1})$; see Figure 12(c). Then Δf_1^{k+1} and Δf_2^{k+1} are identical to those given in the proof of Lemma 2 for the corresponding value of $f(x_{k+1})$. It follows that $\max\{\Delta f_1^{k+1}, \Delta f_2^{k+1}\} \leq \frac{1}{2}\max\{\Delta f_1^k, \Delta f_2^k\}$ and thus $\max\{\Delta f_1^{k+2}, \Delta f_2^{k+2}\} \leq \frac{1}{2}\max\{\Delta f_1^k, \Delta f_2^k\}$.
3. $f^{\mathrm{u}}(x_{k+1}) > f(x_{k+1}) > 0$; see Figure 12(d). The expression for Δf_1^{k+1} is identical to the one given in the proof of Lemma 2, and it follows that $\Delta f_1^{k+1} \leq \frac{1}{2}\Delta f_1^k$. For Δf_2^{k+1} we distinguish two possibilities.
   - $\Delta f_2^{k+1} \leq \frac{1}{2}\Delta f_1^k$, as shown in Figure 12(d). Then $\max\{\Delta f_1^{k+1}, \Delta f_2^{k+1}\} \leq \frac{1}{2}\max\{\Delta f_1^k, \Delta f_2^k\}$ and thus $\max\{\Delta f_1^{k+2}, \Delta f_2^{k+2}\} \leq \frac{1}{2}\max\{\Delta f_1^k, \Delta f_2^k\}$.
   - $\Delta f_2^{k+1} > \frac{1}{2}\Delta f_1^k$, as shown in Figure 13(b). Then also $\Delta f_2^{k+1} > \Delta f_1^{k+1}$. A new point x_{k+2} is now chosen for function evaluation according to (5). For the function value $f(x_{k+2})$, we can also distinguish three possibilities.
     - $f(x_{k+2}) = f^{\mathrm{u}}(x_{k+2})$; see Figure 13(c). As $\Delta f_2^{k+1} > \Delta f_1^{k+1}$, it follows that $\Delta f_2^{k+2} = 0$. Thus,

$$\max\{\Delta f_1^{k+2}, \Delta f_2^{k+2}\} \leq \Delta f_1^{k+1} \leq \frac{1}{2}\max\{\Delta f_1^k, \Delta f_2^k\}.$$

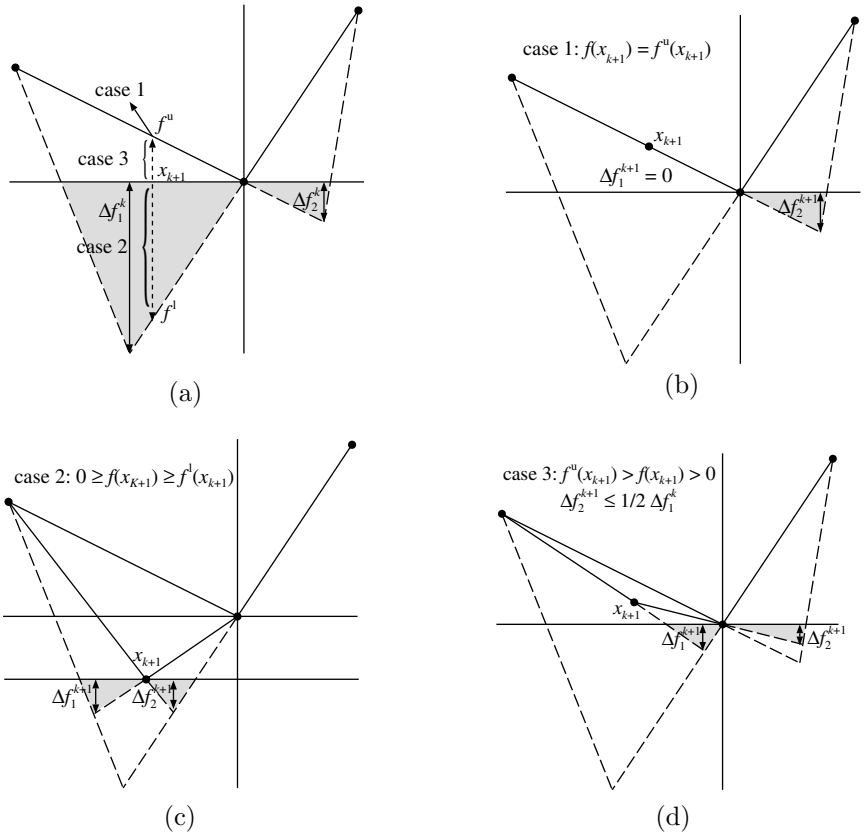


FIG. 12. *Decrease of range of uncertainty when* $\Delta f_1^k \geq \Delta f_2^k > 0$. (a) *Three possibilities are distinguished for the new function evaluation* $f(x_{k+1})$. (b) $f(x_{k+1}) = f^u(x_{k+1})$. (c) $0 \geq f(x_{k+1}) \geq f^l(x_{k+1})$. (d) $f^u(x_{k+1}) > f(x_{k+1}) > 0$.

– $0 \geq f(x_{k+2}) \geq f^l(x_{k+2})$; see Figure 13(d). Then,

$$\max\{\Delta f_1^{k+2}, \Delta f_2^{k+2}\} \leq \frac{1}{2}\Delta f_2^{k+1} \leq \frac{1}{2}\max\{\Delta f_1^k, \Delta f_2^k\}.$$

– $f^u(x_{k+2}) > f(x_{k+2}) > 0$; see Figure 13(e). As $\Delta f_2^{k+1} > \Delta f_1^{k+1}$, it follows that $\Delta f_2^{k+2} \leq \frac{1}{2}\Delta f_2^{k+1} \leq \frac{1}{2}\max\{\Delta f_1^k, \Delta f_2^k\}$. Furthermore, $\Delta f_1^{k+2} \leq \Delta f_1^{k+1} \leq \frac{1}{2}\max\{\Delta f_1^k, \Delta f_2^k\}$. □

COROLLARY 1. *Using the triangle section method, the decrease of the range of uncertainty is exponential in the number of function evaluations.*

Proof. Let the initial range of uncertainty be given by $\max\{\Delta f_1^0, \Delta f_2^0\}$, and let the range of uncertainty after k new function evaluations be given by $\max\{\Delta f_1^k, \Delta f_2^k\}$. Then $\max\{\Delta f_1^k, \Delta f_2^k\} \leq \frac{1}{2}\max\{\Delta f_1^{k-2}, \Delta f_2^{k-2}\} \leq \left(\frac{1}{2}\right)^{\lfloor \frac{k}{2} \rfloor}\max\{\Delta f_1^0, \Delta f_2^0\}$. □

So we have shown that the range of uncertainty is at least halved for each two new function evaluations. However, the average reduction of the range of uncertainty generally is considerably larger, which is backed up by the empirical results that we present in section 5.

3.3. Piecewise-linear functions. We now consider the case in which the function f is also known to be piecewise linear besides convex or concave. An example of such a function is given in section 4.

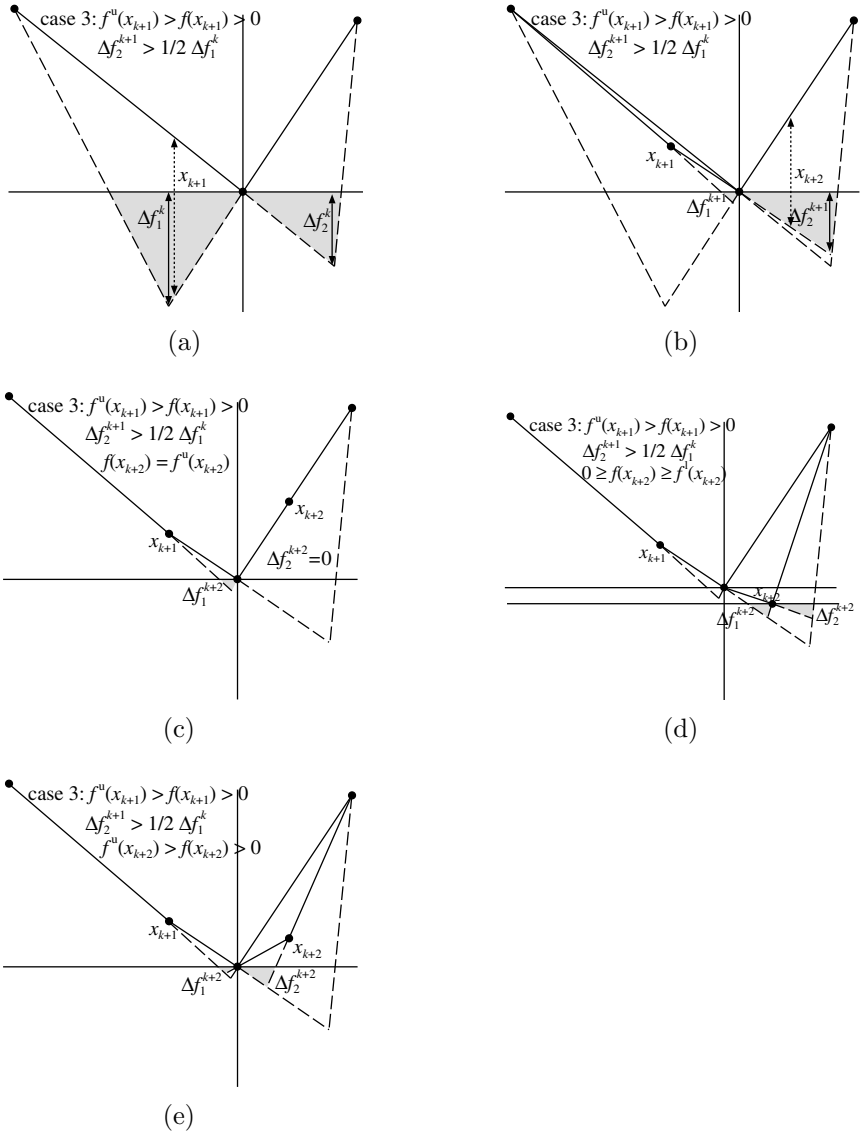


FIG. 13. *Decrease of range of uncertainty when* $\Delta f_1^k \geq \Delta f_2^k > 0$ *and* $\Delta f_2^{k+1} > \frac{1}{2}\Delta f_1^k$. (a) *Initial situation when determining the new function evaluation* $f(x_{k+1})$. (b) $f^u(x_{k+1}) > f(x_{k+1}) > 0$ *and* $\Delta f_2^{k+1} > \frac{1}{2}\Delta f_1^k$. (c) $f(x_{k+2}) = f^u(x_{k+2})$. (d) $0 \geq f(x_{k+2}) \geq f^l(x_{k+2})$. (e) $f^u(x_{k+2}) > f(x_{k+2}) > 0$.

The piecewise-linear property of f can be used as follows to terminate the improved golden section method or the triangle section method with the exact minimum. For a piecewise-linear function the slope of a line segment will be identified as soon as three function evaluations are made of points on the segment. Furthermore, the optimum lies at the intersection of two segments. When these two line segments have been identified, the exact minimum can be obtained by determining the intersection of the two line segments.

When dealing with piecewise-linear functions, we use this fact as follows. When a line segment containing the point with the lowest function evaluation so far has

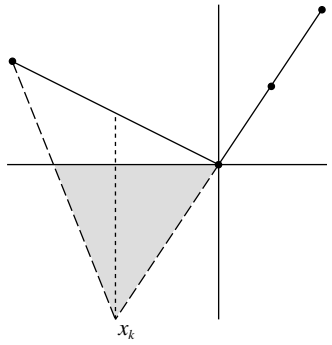


FIG. 14. *A line segment containing the point with the lowest function evaluation so far has been identified by three adjacent function evaluations. The next point x_k that is chosen for function evaluation is the point with the lowest lower bound value.*

been identified, then we evaluate the point x_k with the lowest lower bound value, i.e., $x_k = \arg\min f^{\mathrm{l}}(x)$; see Figure 14 for an example. For the improved golden section method this means that before each new function evaluation we check whether the point with the lowest function evaluation lies on one line segment with the two preceding or succeeding points. For the triangle section method, we check if either $\Delta f_1^k = 0$ or $\Delta f_2^k = 0$ holds. Note that $\Delta f_1^k = 0$ or $\Delta f_2^k = 0$ also holds when there are two points with the lowest function evaluation. If $f(x_k) = f^{\mathrm{l}}(x_k)$, then x_k is the point with the minimum function value, and we can stop. Otherwise, we continue with new function evaluations, also using the latest function evaluation $f(x_k)$.

4. Application to a practical problem: Bandwidth and buffer minimization. In previous sections we described how we can find the optimum for a concave or convex function of one variable with computationally expensive function evaluations. In this section we give a real-life example of such a function, stemming from resource management in an in-home network. For an elaborate description of the problem from which this example originates, we refer to Den Boef, Verhaegh, and Korst [8].

Consider a sender and a receiver of data and a network or data transportation device to which both the sender and receiver are connected. At the connections of the sender and the receiver to the network, buffers are placed. Time is discretized into time units t; for each time unit t, the amount of data supplied at the sender is given by $s(t)$, and the amount of data consumed at the receiver is given by $d(t)$. All data are supplied into the buffer at the sender and are consumed from the buffer at the receiver. Reservations of the buffers and the transportation device are based on the maximum usage during the time horizon. Costs of the buffers are given by c_{s} and c_{r} per unit buffer size, and cost of the transportation device is given by c_{b} per unit transportation capacity.

The problem is to determine reservations b of the transportation device, m_{s} and m_{r} of the buffers, and a feasible transmission schedule of the data given by $x(t)$ for each time unit t such that total costs are minimized. The transmission schedule has to be such that, whenever data is taken from a buffer, it also is available in the buffer (no buffer underflow) and such that, whenever data is put into a buffer, the buffer reservation is not exceeded. Also, the amount transmitted during a time unit may not exceed the transportation capacity reservation. This can be formulated as a linear

program as follows:

$$
\begin{array}{lll}
\text{minimize} & c_{\mathrm{b}}b + c_{\mathrm{s}}m_{\mathrm{s}} + c_{\mathrm{r}}m_{\mathrm{r}} & \\
\text{subject to} & x\,(t) \le b & \forall t, \\
& \sum_{k=1}^{t} s(k) - \sum_{k=1}^{t} x\,(k) \ge 0 & \forall t, \\
& \sum_{k=1}^{t} s(k) - \sum_{k=1}^{t-1} x\,(k) \le m_{\mathrm{s}} & \forall t, \\
& \sum_{k=1}^{t} x\,(k) - \sum_{k=1}^{t} d(k) \ge 0 & \forall t, \\
& \sum_{k=1}^{t} x\,(k) - \sum_{k=1}^{t-1} d(k) \le m_{\mathrm{r}} & \forall t.
\end{array} \tag{9}
$$

In this linear program (LP) all constraints must hold for every time unit t. A typical video stream has a time horizon that is split up into more than 100,000 time units. This leads to an LP that consists of more than $100,000$ variables and $500,000$ constraints, which requires a relatively long calculation time when standard LP-methods are used. However, we can obtain the solution in only a few seconds not by using LP but by using a line search method of this paper together with the following problem-specific method which has complexity $\mathcal{O}(|T|)$, where T denotes the set of time units.

Given a transportation capacity, we can minimize the total buffer costs by first minimizing the buffer with the highest cost coefficient and then minimizing the buffer with the lowest cost coefficient. Minimizing one buffer for a given bandwidth can be done in $\mathcal{O}(|T|)$ time using a specific algorithm [10]. So for a given value of b, optimal values of m_{s} and m_{r} can be determined quickly but without any information on the derivative. This leads to the following reformulation of the problem.

Let f be a function of transportation capacity b with function values that represent the minimum total costs. So, for input b, the function f determines optimal values of m_{s} and m_{r} given the cost coefficients c_{s} and c_{r} and then calculates the total costs $c_{\mathrm{b}}b + c_{\mathrm{s}}m_{\mathrm{s}} + c_{\mathrm{r}}m_{\mathrm{r}}$ which are returned as output. The problem now is to find the minimum of the function f. Since the original problem is an LP-problem with b appearing in the right-hand sides of the constraints, f is a continuous, piecewise-linear, convex function. So, the method described in this paper can be used to find an optimum. In section 5, we show some results of the improved golden section method and the triangle section method applied to this problem.

5. Numerical test results. In this section we present numerical test results for the improved golden section method and the triangle section method. These results are obtained by using these methods on the bandwidth-buffer application described in section 4 and on numerous mathematical functions, which can be divided into two types. Functions of type 1 are polynomial functions, given by $f(x) = a(x-b)^{2c}$, with $a = 0.5, 1, 1.5, \ldots, 9.5, 10$, $b = 1, 2, \ldots, 10$, and $c = 1, 2, \ldots, 5$. This gives a total of 1000 functions of type 1. Note that type 1 functions are symmetrical with a steep grade. A function f of type 2 is given by $f(x) = a\mathrm{e}^{b(x-c)} - dx$, with $a = 1, 2, \ldots, 10$, $b = 1, 2, \ldots, 5$, $c = -5, -4, \ldots, 5$, and $d = 0.01, 0.05, 0.25, 1.25, 6.25, 31.25, 156.25, 781.25, 3906.25$. This gives a total of 4950 functions of type 2. Note that type 2 functions are nonsymmetrical and have a steep grade on one side of the minimum and a weak grade on the other side of the minimum. For both functions of type 1 and functions of type 2 the objective was to find the minimum on the interval $[-10, 10]$. Finally, for the bandwidth-buffer application discussed in the previous section, we used 16 different video traces, each combined with a number of different cost coefficients c_{b}, c_{s}, and c_{r}, which resulted in 283 problem instances.

In Table 5.1 results are given for comparing the regular golden section, improved golden section, and triangle section methods using (a) the 1000 functions of type 1, (b) the 4950 functions of type 2, and (c) the 283 application instances. The results

TABLE 5.1

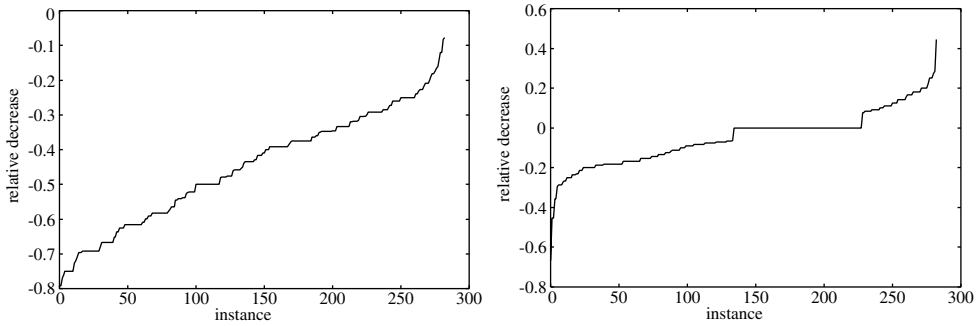
Results after 5–10 *function evaluations for the golden section method (gs), the improved golden section method (igs), and the triangle section method (ts).* (a), (b), *and* (c) *give the results for the functions of types* 1 *and* 2, *and for the bandwidth-buffer application, respectively. In each table, the leftmost column gives the number of function evaluations. The next three columns give for each method and number of function evaluations the average over all functions or bandwidth-buffer instances of the difference between the function value of the best found solution and the optimal function value. The three rightmost columns give for each method the average over all functions or bandwidth-buffer instances of the resulting interval of uncertainty after the given number of function evaluations. Finally, improved golden section and triangle section methods were stopped when for a function the range of uncertainty was less than* 0.01. *When this happened for a function, the last calculated result of the best found solution and the interval of uncertainty was also used as a result for the higher number of function evaluations in this table. For example, if the triangle section method was stopped after eight function evaluations with best found solution value* 2.5 *and interval of uncertainty size* 0.5, *then these values were also used for nine and ten function evaluations to compute the results in this table. This explains why, e.g., the average size of the interval of uncertainty for type* 1 *functions using the triangle section method is the largest, compared to regular and improved golden section methods, for ten function evaluations, while it is the smallest for fewer function evaluations.*

(a) Type 1	Avg. deviation from optimum			Avg. int. of uncertainty		
# func.eval.	gs	igs	ts	gs	igs	ts
5	7,126.103	8.087	2.054	6.111	5.733	5.638
6	3.649	3.137	0.163	3.708	3.404	2.993
7	0.200	0.192	0.008	2.265	2.076	1.678
8	0.068	0.041	0.003	1.390	1.264	1.069
9	0.030	0.013	0.001	0.855	0.770	0.799
10	0.005	0.002	0.000	0.527	0.470	0.765

(b) Type 2	Avg. deviation from optimum			Avg. int. of uncertainty		
# func.eval.	gs	igs	ts	gs	igs	ts
5	605.264	531.365	727.436	6.912	4.606	5.098
6	283.096	229.274	272.258	4.262	2.408	2.736
7	145.214	87.939	89.639	2.631	1.368	1.528
8	68.755	31.416	32.209	1.625	0.784	0.906
9	32.362	15.685	16.408	1.004	0.447	0.575
10	17.983	12.590	12.595	0.620	0.255	0.409

(c) Bw.-buf.	Avg. deviation from optimum			Avg. int. of uncertainty		
# func.eval.	gs	igs	ts	gs	igs	ts
5	112,082	81,744	70,924	5,079	3,990	3,362
6	40,726	28,890	23,090	3,100	1,803	1,831
7	19,950	11,894	10,123	1,900	839	1,017
8	11,651	7,053	5,585	1,168	391	610
9	7,301	3,857	2,902	719	182	407
10	4,255	1,945	1,712	444	78	325

consist of the average difference between the best found solution and the optimal solution and the average size of the interval of uncertainty after a given number of function evaluations, with both averages over all functions or instances. These results show that for functions of type 1 and bandwidth-buffer instances, the triangle section method on average gets closest to the optimal function value. For functions of type 2, however, it is outperformed by the improved golden section method, possibly due to the largely asymmetrical shape of these functions. Regarding the interval of uncertainty, we notice that the improved golden section method improves upon the results of the regular golden section method especially for functions of type 2 and for the bandwidth-buffer instances. For functions of type 1, it also returns better results than the regular golden section method, but here the improvement is not so



(a) (b)

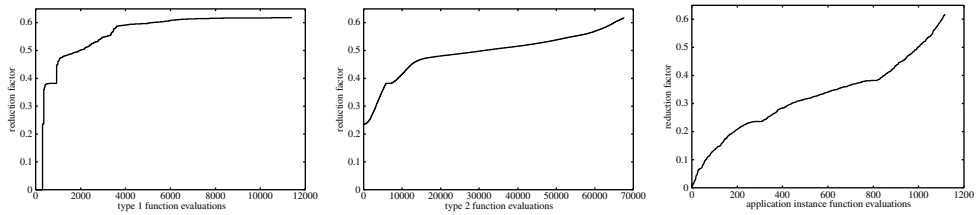
FIG. 15. *In these two graphs the number of function evaluations that two of the three methods require are compared.* (a) *compares the improved golden section method with the regular golden section method for the* 283 *bandwidth-buffer instances. Both methods were stopped when the interval of uncertainty was less than* 1. *The graph gives for all* 283 *instances the relative decrease in the number of function evaluations required by the improved golden section method compared to the regular golden section method, i.e., for each application instance* $(igs - gs)/gs$ *is given, where* igs *and* gs *denote the number of function evaluations required for the improved golden section and regular golden section methods, respectively. The instances are sorted on the horizontal axis in the order of nondecreasing relative decrease.* (b) *compares the triangle section method to the improved golden section method, both using the procedure that uses piecewise linearity to determine the exact optimum. The graph gives for all* 283 *instances the relative decrease in the number of function evaluations required by the triangle section method compared to the improved golden section method, i.e., for each bandwidth-buffer instance* $(ts - igs)/igs$ *is given, where* ts *and* igs *denote the number of function evaluations required for the triangle section and improved golden section methods, respectively.*

impressive. This can be explained by the relatively steep slopes of the functions of type 1 surrounding both sides of the optimum.

Figure 15(a) compares the improved golden section method with the regular golden section method using the bandwidth-buffer instances and as stopping criterion the size of interval of uncertainty being less than 1. It shows that the reduction in the number of function evaluations can be as high as 80%, and on average around 40%. Figure 15(b) compares the triangle section method with the improved golden section method using the application instances and the piecewise-linear property, thus obtaining the exact optimum. For about one-half of the instances the triangle section method requires fewer function evaluations than the improved golden section method. However, for about one-quarter of the instances it requires more function evaluations. Still, the average number of function evaluations is lower for the triangle section method than for the improved golden section method.

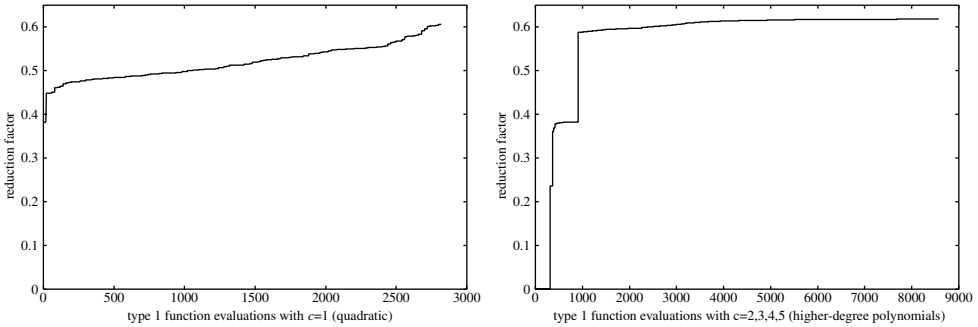
Figure 16 shows the reduction factors of the interval of uncertainty after a function evaluation, which were observed when applying the improved golden section method on all functions and instances. It shows that the reduction factor is often close to the golden section (≈ 0.618) for functions of type 1. However, for functions of type 2 and bandwidth-buffer instances, the reduction is much more significant, i.e., we get much smaller intervals of uncertainty.

Figure 17 shows again the reduction factors of the interval of uncertainty using the improved golden section method for functions of type 1, but now split into quadratic functions ($c = 1$) and functions of higher degree ($c \geq 2$). It shows that the reduction factors for quadratic functions are distributed evenly between 0.45 and 0.6 in contrast with the reduction factors for higher-degree functions, of which approximately 90%



(a) (b) (c)

FIG. 16. *These graphs show on the vertical axis the observed reduction factors of the interval of uncertainty after a function evaluation using the improved golden section method, where the function evaluations are sorted in the order of increasing reduction factor. Note that the number of performed function evaluations may vary for different functions and bandwidth-buffer instances. Also note that a smaller reduction factor leads to a smaller new interval of uncertainty and thus is better.* (a) *gives the reduction factors for all* 11,397 *function evaluations that were performed for the* 1,000 *type* 1 *functions,* (b) *for all* 67,570 *evaluations of the* 4,950 *type* 2 *functions, and* (c) *for all* 1,119 *evaluations of the* 283 *application instances.*



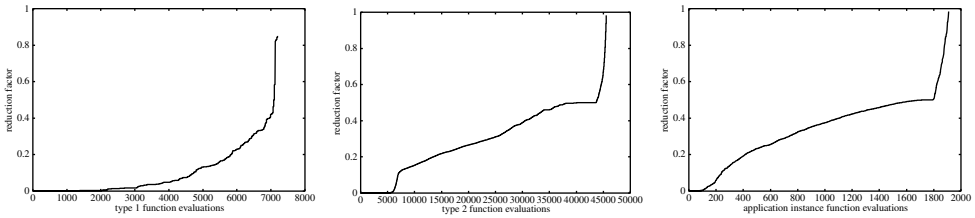
(a) (b)

FIG. 17. *These graphs show on the vertical axis the observed reduction factors of the interval of uncertainty after a function evaluation using the improved golden section method, where the function evaluations are again sorted in the order of increasing reduction factor.* (a) *gives the reduction factors for all* 2,820 *function evaluations that were performed for the* 200 *quadratic type* 1 *functions,* (b) *for all* 8,577 *function evaluations of the* 800 *other type* 1 *functions.*

are close to 0.6. As the gradient of functions of higher degree at the evaluated points is larger than the gradient of quadratic functions, the lower bounds for the functions of higher degree have a steeper slope and thus lead to a smaller reduction of the interval of uncertainty.

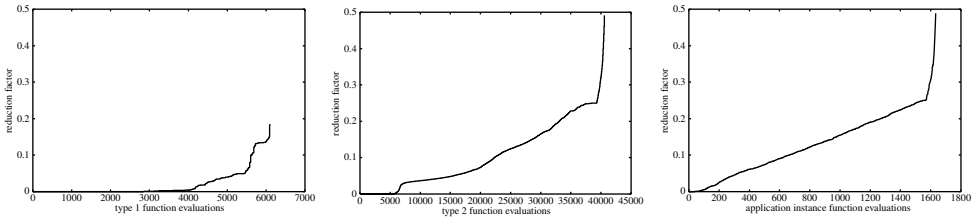
Figure 18 shows the reduction factors of the range of uncertainty after a single function evaluation, which were observed when applying the triangle section method on all functions and instances. It shows that more than 90% of the recorded reduction factors are spread out over the interval $[0, 0.5]$, and only a small fraction of the recorded reduction factors are between 0.5 and 1. Figure 19 shows the reduction factors of the range of uncertainty after two consecutive function evaluations; cf. Theorem 2. It shows a similar distribution, with most of the observed reduction factors between 0 and 0.25 and only a small fraction between 0.25 and 0.5.

6. Conclusion. In this paper we have considered the line search problem for convex functions. We have shown how the convexity property can be used to obtain upper and lower bounds on the function using the performed function evaluations. For some well-known line search methods we have shown, using these upper and



(a) (b) (c)

FIG. 18. *These graphs show on the vertical axis the observed reduction factors of the range of uncertainty after a single function evaluation using the triangle section method, where the evaluations are sorted in the order of increasing reduction factor. These reduction factors were observed after at least three functions evaluations were made for a specific function or bandwidth-buffer instance. Note that also for the triangle section method the number of performed function evaluations can vary between different functions and bandwidth-buffer instances.* (a) *gives the reduction factors for all* 7,214 *evaluations that were performed for the* 1,000 *type* 1 *functions,* (b) *for all* 45,525 *evaluations of the* 4,950 *type* 2 *functions, and* (c) *for all* 1,911 *evaluations of the* 283 *bandwidth-buffer instances.*



(a) (b) (c)

FIG. 19. *These graphs show on the vertical axis the observed reduction factors of the range of uncertainty after two consecutive function evaluations using the triangle section method, where the evaluations are sorted in the order of increasing reduction factor. These reduction factors were observed after at least three functions evaluations were made for a specific function or bandwidth-buffer instance.* (a) *gives the reduction factors for all* 6,114 *pairs of consecutive function evaluations that were performed for the* 1,000 *type* 1 *functions,* (b) *for all* 40,576 *pairs of consecutive evaluations of the* 4,950 *type* 2 *functions, and* (c) *for all* 1,634 *pairs of consecutive evaluations of the* 283 *bandwidth-buffer instances.*

lower bounds, that they may propose a candidate which is not optimal. We have presented two new line search methods which use the convexity property. The first method, the improved golden section method, uses the upper and lower bounds to improve upon the regular golden section method and always proposes a candidate which can be optimal. The second method, the triangle section method, focuses on minimizing the interval for possible objective values, the range of uncertainty, and we have shown that it at least halves the range of uncertainty after every two function evaluations.

Both methods were tested using a real-life example and two classes of convex functions. It was shown that the new methods give better approximations of the optimum than the regular golden section method after a fixed number of function evaluations. This also translated into a sometimes heavily reduced number of function evaluations that was required to obtain the optimum. A direct comparison of the new methods did not show a clear winner; depending on the instance either the improved golden section or the triangle section method gave the best results.

There are several possibilities for future research in line search methods for convex functions. The upper and lower bounds based on the convexity property can be used

to adapt other well-known line search methods. They can also be used to try to estimate the complete function as efficiently as possible instead of only the optimum. When information on the derivative in an evaluated point is known, it can be used to further improve the interval and range of uncertainty. This will also improve the performance of the methods presented in this paper. Finally, it would be interesting to see how the work presented in this paper extends to multivariate, convex functions.

Appendix A. Height of a triangle in the area of uncertainty. For the explanation of the triangle section method to obtain a guaranteed range reduction, the heights of the triangles in the area of uncertainty, Δf_1^k and Δf_2^k, need to be determined. The expressions of Δf_1^k and Δf_2^k in points L, M, and U with their function evaluations $f(L), f(M)$, and $f(U)$ and the interval of uncertainty $[L', U']$ can be obtained by determining the intersection of two appropriate lines and subtracting this from the smallest function evaluation so far. We first give a general expression of the intersection of two lines both determined by two points. Then we indicate for (3), (4), (6), (7), and (8) how they can be obtained.

Let $(a, f(a))$ and $(b, f(b))$ define a line l_1, and let $(c, f(c))$ and $(d, f(d))$ define a line l_2. So, $l_1(x)$ is given by

$$\frac{f(b) - f(a)}{b - a}x + \frac{f(a)b - f(b)a}{b - a},$$

and $l_2(x)$ is given by

$$\frac{f(d) - f(c)}{d - c}x + \frac{f(c)d - f(d)c}{d - c}.$$

Let line l_1 and line l_2 intersect at $(y, f(y))$. Then using the above line equations with $x = y$, we derive from $l_1(y) = l_2(y)$ that

$$y = \frac{\frac{f(c)d - f(d)c}{d - c} - \frac{f(a)b - f(b)a}{b - a}}{\frac{f(b) - f(a)}{b - a} - \frac{f(d) - f(c)}{d - c}}.$$

Simplifying the above equation gives

$$y = \frac{(f(c)d - f(d)c)(b - a) - (f(a)b - f(b)a)(d - c)}{(f(b) - f(a))(d - c) - (f(d) - f(c))(b - a)}.$$

Substituting the above equation for y into $l_1(y)$ or $l_2(y)$ gives for $f(y)$

$$f(y) = \frac{(f(c)d - f(d)c)(f(b) - f(a)) - (f(a)b - f(b)a)(f(d) - f(c))}{(f(b) - f(a))(d - c) - (f(d) - f(c))(b - a)}.$$

Equation (3):
$(a, f(a)) = (L, f(L))$, $(b, f(b)) = (L', f(M))$, $(c, f(c)) = (M, f(M))$, and $(d, f(d)) = (U, f(U))$.
$\Delta f_1^k = f(M) - f(y)$.

Equation (4):
$(a, f(a)) = (L, f(L))$, $(b, f(b)) = (M, f(M))$, $(c, f(c)) = (U', f(M))$, and $(d, f(d)) = (U, f(U))$.
$\Delta f_1^k = f(M) - f(y)$.

Equation (6):
$(a, f(a)) = (A, B)$, $(b, f(b)) = (-\frac{1}{2}, E)$, $(c, f(c)) = (0, 0)$, and $(d, f(d)) = (D, f(D))$.
$\Delta f_1^{k+1} = 0 - f(y)$.

Equation (7):
$(a, f(a)) = (A, B)$, $(b, f(b)) = (-1, 0)$, $(c, f(c)) = (-\frac{1}{2}, E)$, and $(d, f(d)) = (0, 0)$.
$\Delta f_1^{k+1} = E - f(y)$.

Equation (8):
$(a, f(a)) = (A, B)$, $(b, f(b)) = (-\frac{1}{2}, E)$, $(c, f(c)) = (0, 0)$, and $(d, f(d)) = (D, f(D))$.
$\Delta f_1^{k+1} = E - f(y)$.

REFERENCES

[1] K. A. Ariyawansa and W. L. Tabor, *A note on line search termination criteria for collinear scaling algorithms*, Computing, 70 (2003), pp. 25–39.

[2] B. W. Bader and R. B. Schnabel, *Curvilinear linesearch for tensor methods*, SIAM J. Sci. Comput., 25 (2003), pp. 604–622.

[3] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms*, 2nd ed., Wiley, New York, 1993.

[4] M. Bendaya, *Line search techniques for the logarithmic barrier function in quadratic-programming*, J. Oper. Res. Soc., 46 (1995), pp. 332–338.

[5] H. Y. Benson, D. F. Shanno, and R. J. Vanderbei, *Interior-point methods for nonconvex nonlinear programming: Jamming and numerical testing*, Math. Program., 99 (2004), pp. 35–48.

[6] A. R. Conn, N. I. M. Gould, and P. L. Toint, *Trust-Region Methods*, Vol. 1, MPS/SIAM Ser. Optim. 1, SIAM, Philadelphia, 2000.

[7] Y. H. Dai, *On the nonmonotone line search*, J. Optim. Theory Appl., 112 (2002), pp. 315–330.

[8] E. den Boef, W. F. J. Verhaegh, and J. Korst, *Smoothing streams in an in-home digital network: Optimization of bus and buffer usage*, Telecommunication Syst., 23 (2003), pp. 273–295.

[9] D. den Hertog, *Interior Point Approach to Linear, Quadratic and Convex Programming: Algorithms and Complexity*, Mathematics and its Applications 277, Kluwer Academic Publishers, Dordrecht, 1994.

[10] W. Feng, *Rate-constrained bandwidth smoothing for the delivery of stored video*, in Proceedings of the SPIE Multimedia Networking and Computing Conference, Society of Photo-optical Instrumentation Engineers, 1997, pp. 316–327.

[11] P. E. Gill, W. Murray, and M. H. Wright, *Practical Optimization*, 7th ed., Academic Press, San Diego, CA, 1988.

[12] J. Guérin, P. Marcotte, and G. Savard, *An optimal adaptive algorithm for the approximation of concave functions*, Math. Program., 107 (2006), pp. 357–366.

[13] J.-B. Hiriart-Urruty and C. Lemarechal, *Convex Analysis and Minimization Algorithms, Vols. 1 and 2*, Springer, Berlin, 1993.

[14] D. H. Li and M. Fukushima, *A derivative-free line search and global convergence of Broyden-like method for nonlinear equations*, Optim. Methods Softw., 13 (2000), pp. 181–201.

[15] A. Melman, *A new linesearch method for quadratically constrained convex-programming*, Oper. Res. Lett., 16 (1994), pp. 67–77.

[16] A. Melman, *A linesearch procedure in barrier methods for some convex programming problems*, SIAM J. Optim., 6 (1996), pp. 283–298.

[17] W. Murray and M. H. Wright, *Line search procedures for the logarithmic barrier function*, SIAM J. Optim., 4 (1994), pp. 229–246.

[18] F. A. Potra and Y. Shi, *Efficient line search algorithm for unconstrained optimization*, J. Optim. Theory Appl., 85 (1995), pp. 677–704.

[19] C. Roos, T. Terlaky, and J.-Ph. Vial, *Theory and Algorithms for Linear Optimization: An Interior Point Approach*, Wiley, Chichester, 1997.

[20] P. L. Toint, *An assessment of nonmonotone linesearch techniques for unconstrained optimization*, SIAM J. Sci. Comput., 17 (1996), pp. 725–739.

[21] D. X. Xie and T. Schlick, *A more lenient stopping rule for line search algorithms*, Optim. Methods Softw., 17 (2002), pp. 683–700.

SIAM J. OPTIM.
Vol. 18, No. 1, pp. 364–378

NEW KORKIN–ZOLOTAREV INEQUALITIES*

R. A. PENDAVINGH† AND S. H. M. VAN ZWAM†

Abstract. Korkin and Zolotarev showed that if

$$\sum_i A_i \Big(x_i - \sum_{j>i} \alpha_{ij} x_j \Big)^2$$

is the Lagrange expansion of a Korkin–Zolotarev (KZ-) reduced positive definite quadratic form, then $A_{i+1} \geq \frac{3}{4} A_i$ and $A_{i+2} \geq \frac{2}{3} A_i$. They showed that the implied bound $A_5 \geq \frac{4}{9} A_1$ is not attained by any KZ-reduced form. We propose a method to optimize numerically over the set of Lagrange expansions of KZ-reduced quadratic forms using a semidefinite relaxation combined with a branch and bound process. We use a rounding technique to derive exact results from the numerical data. Applying these methods, we prove several new linear inequalities on the A_i of any KZ-reduced form, one of them being $A_{i+4} \geq (\frac{15}{32} - 2 \cdot 10^{-5}) A_i$. We also give a form with $A_5 = \frac{15}{32} A_1$. These new inequalities are then used to study the cone of outer coefficients of KZ-reduced forms, to find bounds on Hermite's constant, and to give better estimates on the quality of k-block KZ-reduced lattice bases.

Key words. lattice, quadratic form, semidefinite programming, Korkin–Zolotarev reduction, Hermite's constant, sphere packing

AMS subject classifications. 11H55, 52C17, 90C22, 11H50

DOI. 10.1137/060658795

1. Preliminaries and overview.

The *Geometry of Numbers* is a field of mathematics initiated and named by Minkowski. The main objects studied are *lattices*, discrete subgroups of $\mathbb{R}^n$. Good introductions to the subject are the book by Cassels [2] and the excellent survey paper by Ryškov and Baranovskiĭ [13]. Typical problems are the search for a shortest vector within a given lattice and the search for a lattice with a dense *sphere packing*. Hermite's constant γ_n is a measure for the density of the densest lattice sphere packing in dimension n. This constant has been determined exactly for $n \leq 8$ and $n = 24$. Since Blichfeldt [1] determined γ_n for $n = 6, 7, 8$, no further low-dimensional cases have been computed. For example, the best known bounds for $n = 9$ are $512 \leq \gamma_9^9 < 913$, where the lower bound is the density of a specific lattice (see, for example, [4]), and the upper bound is the Cohn–Elkies bound [3].

Most of the early research in this subject was not in terms of lattices but in terms of quadratic forms. This approach proved very useful for our research, so in all but the last section we will talk exclusively about positive definite quadratic forms.

An n-ary positive definite quadratic form q can be written uniquely as

$$q(x_1, \ldots, x_n) = \sum_{i=1}^{n} A_i \Big(x_i - \sum_{j>i} \alpha_{ij} x_j \Big)^2. \tag{1.1}$$

This is the *Lagrange expansion* of q; the numbers A_i are the *outer coefficients* and

*Received by the editors May 3, 2006; accepted for publication (in revised form) December 6, 2006; published electronically May 1, 2007. Parts of this research appeared previously in [16].
http://www.siam.org/journals/siopt/18-1/65879.html
†Technische Universiteit Eindhoven, P.O. Box 513, 5600 MB Eindhoven, The Netherlands (rudi@win.tue.nl, svzwam@win.tue.nl).

the α_{ij} the *inner coefficients*. We write

$$q_k(x_k,\dots,x_n) := \sum_{i=k}^{n} A_i\Big(x_i - \sum_{j>i}\alpha_{ij}x_j\Big)^2. \tag{1.2}$$

A positive definite quadratic form q in n variables with Lagrange expansion (1.1) is *Korkin–Zolotarev (KZ-)*[1] *reduced* if

$$|\alpha_{ij}| \le \frac{1}{2} \text{ for all } i,j, \text{ and } \alpha_{i,i+1} \ge 0 \text{ for all } i; \tag{S}$$

and

$$A_k \le q_k(x) \text{ for all nonzero } x \in \mathbb{Z}^{n-k+1}, k = 1,\dots,n-1. \tag{M}$$

We say that two forms q, q' are *equivalent* if there is a unimodular matrix U, i.e., $U \in GL_n(\mathbb{Z})$, such that $q'(x) = q(Ux)$. It can be shown that any form is equivalent to a KZ-reduced form (see, for example, [13]).

Korkin and Zolotarev proved that the outer coefficients of a KZ-reduced form satisfy $A_2 \ge \frac{3}{4}A_1$ (the *first KZ-inequality*) and $A_3 \ge \frac{2}{3}A_1$ (the *second KZ-inequality*) [7]. If q is KZ-reduced, then so is the quadratic form q_k for $k \ge 1$, and hence the inequalities

$$A_{k+1} \ge \frac{3}{4}A_k \text{ and } A_{k+2} \ge \frac{2}{3}A_k,\ k = 1,2,\dots, \tag{1.3}$$

hold for the outer coefficients of any KZ-reduced form.

For each $n \in \mathbb{N}$, *Hermite's constant* is defined as

$$\gamma_n := \max\left\{ \frac{m(q)}{\det(q)^{\frac{1}{n}}} \mid q \text{ is an } n\text{-ary positive definite quadratic form}\right\}, \tag{1.4}$$

where $m(q) := \min\{q(x) \mid x \in \mathbb{Z}^n, x \ne 0\}$ is the *minimum* of the form q and $\det(q) := \det(Q)$, where Q is the symmetric matrix such that $q(x) = x^tQx$. Equivalent forms have the same minimum and the same determinant, so we may as well restrict the feasible set of (1.4) to KZ-reduced forms. Also, if $A_1,\dots,A_n$ are the outer coefficients of a form q, then $\det(q) = \prod_i A_i$, and if q is KZ-reduced, then $m(q) = f(1,0,\dots,0) = A_1$. Hence

$$\gamma_n^n = \max\left\{ \frac{A_1^n}{\prod_i A_i} \mid (A_1,\dots,A_n) = A(q) \text{ for some KZ-reduced form } q\right\}, \tag{1.5}$$

where $A(q) := (A_1,\dots,A_n)$ denotes the sequence of outer coefficients of the quadratic form q. Using (1.3), this implies the bound

$$\gamma_n^n \le \max\left\{ \frac{A_1^n}{\prod_{i=1}^n A_i} \mid A_{i+1} \ge \frac{3}{4}A_i, A_{i+2} \ge \frac{2}{3}A_i, A_1 = 1\right\}, \tag{1.6}$$

[1] In the literature one encounters several different ways of writing the names of Korkin and Zolotarev. We decided to follow some of the more recent publications (notably [5]) and have kept the original spelling in the references to facilitate the search for these papers.

which is tight for $n = 2, 3, 4$, as was shown in [6, 7]. In the right-hand side of (1.6) we have removed the scale invariance by requiring $A_1 = 1$. The right-hand side is equal to the inverse of

$$\min\left\{\prod_{i=1}^{n} A_i \mid A_{i+1} \geq \frac{3}{4}A_i, A_{i+2} \geq \frac{2}{3}A_i, A_1 = 1\right\}, \tag{1.7}$$

which is the minimum of a concave function on a polyhedron. It is a basic fact of convex optimization that this minimum is attained at one of the vertices. Enumerating all vertices now suffices to determine the bound.

The proof of the first KZ-inequality is elementary. The proof of the second KZ-inequality also uses elementary techniques but is already quite involved. To prove an upper bound on γ_5, Korkin and Zolotarev developed other techniques [8]: they characterized the local optima of the objective function of (1.4), which enabled them to enumerate all local optima for $n = 5$. This line of investigation has been continued and is still actively pursued [10].

In this paper, we return to the first approach and focus on the feasible set of (1.5). We develop a method to prove linear inequalities that hold for the outer coefficients of KZ-reduced forms. Our method is numerical and uses recently developed polynomial optimization techniques. We apply our method in particular to forms in five variables and obtain inequalities (Theorems 6.1 and 6.2) that imply, through (1.5), an upper bound on γ_n that is very close to the known value for $n = 5, 6, 7, 8$.

The structure of the paper is as follows. In the next section, we give preliminaries on KZ-reduced forms. In particular, we describe results of Novikova [11] that imply that the set of KZ-reduced forms can be defined by finitely many polynomial inequalities. Proving that a linear inequality on the outer coefficients holds for KZ-reduced forms thus amounts to minimizing the value of a polynomial under finitely many polynomial constraints.

Through recent developments in convex optimization it is possible to find lower bounds for such polynomial optimization problems using semidefinite optimization methods. We describe such a semidefinite relaxation in section 3.

We improve on the lower bound that results from simply computing the semidefinite relaxation by performing a *branch and bound* procedure, which is familiar from integer programming. By splitting the semialgebraic set over which we are optimizing we obtain a number of problems on smaller sets. The relaxation for each of these smaller problems is stronger than the original relaxation and will yield a higher lower bound. Then the smallest of these lower bounds is again a lower bound for the original problem. The branch and bound procedure is described in section 4.

Although we use a numerical method, our final results are exact in the sense that their validity does not depend on the accuracy with which the floating point computations were performed. Each of the many lower bounds we have computed is determined by a convex optimization problem which has a well-defined convex dual. By rounding each optimal dual solution to a nearby rational and feasible solution, an exact lower bound is obtained. Its validity can be verified independently, using only elementary rational arithmetic. The rounding method is described in section 5.

In section 6 we derive, using these tools, several new linear inequalities on the outer coefficients of KZ-reduced forms. We study the relation between these inequalities and the cone of outer coefficients of KZ-reduced forms. The most striking result is that only three of these new inequalities suffice to give very good bounds on Hermite's constant up to dimension 8.

Finally, in section 7 we show how our new inequalities on the outer coefficients lead to better quality estimates for the block KZ-reduction algorithm.

The implementation and verification of our numerical method is worked out in detail in [17].

2. A finite characterization of KZ-reduced forms. A positive definite quadratic form q of two or more variables is KZ-reduced if (**S**) holds, if q_2 is KZ-reduced, and if

$$A_1 \le q(x) \text{ for all nonzero } x \in \mathbb{Z}^n. \tag{2.1}$$

In [11], Novikova stated the following.

THEOREM 2.1. *For each $n \ge 2$, there is a finite set $X_n \subseteq \mathbb{Z}^n$ such that an n-ary form with Lagrange expansion* (1.1) *is KZ-reduced if and only if q_2 is KZ-reduced,* (**S**) *holds, and*

$$A_1 \le q(x) \text{ for all } x \in X_n. \tag{2.2}$$

The proof boils down to the fact that if q_2 is KZ-reduced, $q(0,1,0,\dots,0) \ge A_1$, and (**S**) holds, then $q(x) \ge A_1$ is implied for all but finitely many $x \in \mathbb{Z}^n$. This argument yields highly redundant sets X_n. But the theorem implies the existence of a unique irredundant set X_n, which we will denote by X_n^*. In [11], Novikova gives finite sets X_n for $n \le 8$ and claims irredundancy of these sets for $n \le 5$. It is unfortunate that the proofs were omitted from her paper, as it appears to be a significant challenge to determine these irredundant sets. We were only able to verify her claims for $n \le 4$. For $n \in \{5,6\}$ we find sufficient sets that are slightly larger, and for larger n the sets we compute are much smaller [16]. We have proven necessity for all vectors up to dimension 4 and all of Novikova's vectors in dimension 5.

It is easy to see that $X_n^* = \{x \in \mathbb{Z}^n \mid (x,0) \in X_{n+1}^*\}$ for any $n \ge 2$. Let $\bar{X} := \{(x,0) \mid x \in X\}$. One has

$$X_2^* = \left\{ \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\}, \tag{2.3}$$

$$X_3^* \setminus \bar{X}_2^* = \left\{ \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right\}. \tag{2.4}$$

Moreover, $X_4^* \setminus \bar{X}_3^*$ is a set of 12 vectors, and according to Novikova $X_5^* \setminus \bar{X}_4^*$ is a set of 52 vectors [11].

Using Theorem 2.1 we find that in the definition of KZ-reducedness, the requirement (**M**) is equivalent to

$$A_k \le q_k(x) \text{ for all } x \in X_{n-k+1}^*, k = 1, \dots, n-1. \tag{**N**}$$

Thus $(A_1, \dots, A_n, \alpha_{12}, \dots, \alpha_{n-1,n})$ are the outer and inner coefficients of a KZ-reduced form if and only if they satisfy finitely many linear inequalities (**S**) and finitely many cubic inequalities (**N**). The number of inequalities of the second kind seems to grow much faster than that of the first kind as n increases.

It is possible to characterize the KZ-reduced forms using only linear and quadratic inequalities by using a different parametrization of the set of quadratic forms. Let Q

be a positive definite $n \times n$ matrix and let $q(x) := x^t Q x$. Then the Lagrange expansion (1.1) yields a decomposition

$$Q = \sum_{i=1}^{n} a_i^t a_i = C^t C, \tag{2.5}$$

where

$$a_i = \sqrt{A_i}(0, \ldots, 0, 1, -\alpha_{i,i+1}, \ldots, -\alpha_{in}) \tag{2.6}$$

is a row vector for $i = 1, \ldots, n$ and C is the matrix whose ith row is a_i.

Thus C is upper triangular, and $Q = C^t C$ is the Cholesky decomposition of Q.

Let $S^i := [0, \frac{1}{2}] \times [-\frac{1}{2}, \frac{1}{2}]^{n-i-1}$. Then

$$\left\{ \sqrt{A_i}(0, \ldots, 0, 1, -\alpha_{i,i+1}, \ldots, -\alpha_{in}) \mid A_i \geq 0, (\alpha_{i,i+1}, \ldots, \alpha_{in}) \in S^i \right\} \tag{2.7}$$

is a polyhedral cone, so there is a finite set of column vectors, which we call D_i, such that (2.7) equals

$$\{a \in \mathbb{R}^n \mid ad \geq 0 \text{ for all } d \in D_i\}. \tag{2.8}$$

For $x \in \mathbb{Z}^m$, $m \leq n$, we write $\widetilde{x} := (0, \ldots, 0, x_1, \ldots, x_m) \in \mathbb{Z}^n$. Now $q(x) = x^t Q x$ is KZ-reduced if and only if there are row vectors $a_i \in \mathbb{R}^n$ such that $Q = \sum_i a_i^t a_i$ and

(**S'**) $$a_k d \geq 0 \text{ for all } d \in D_k \text{ for } k = 1, \ldots, n;$$

and

(**N'**) $$\sum_{i=k}^{n} (a_i \widetilde{x})^2 \geq a_{kk}^2 \text{ for all } x \in X^*_{n-k+1}, k = 1, \ldots, n-1.$$

3. A semidefinite relaxation. The characterizations above describe the coefficient domain of KZ-reduced forms as a semialgebraic set. There is by now a standard machinery for constructing semidefinite relaxations for the problem of minimizing a polynomial over a semialgebraic set; see [9, 12]. We describe a semidefinite formulation that has the virtue of yielding a reasonable lower bound while using only a moderate number of variables and constraints.

THEOREM 3.1. *Let Q be an $n \times n$ positive definite matrix and let $q(x) = x^t Q x$. Then q is KZ-reduced if and only if there are $n \times n$ matrices $B^1, \ldots, B^n$ such that $Q = B^1 + \cdots + B^n$ and*

(**r**) $$B^k \text{ has rank 1 for } k = 1, \ldots, n;$$

(**p**) $$B^k \text{ is positive semidefinite for } k = 1, \ldots, n;$$

(**s**) $$d_1^t B^k d_2 \geq 0 \text{ for all } d_1, d_2 \in D_k, \text{ for } k = 1, \ldots, n; \text{ and}$$

(**n**) $$\sum_{i=k}^{n} \widetilde{x}^t B^i \widetilde{x} \geq B^k_{kk} \text{ for all } x \in X^*_{n-k+1}, \text{ for } k = 1, \ldots, n-1.$$

Proof. To see necessity, let q be KZ-reduced and let A_i, α_{ij} be its outer and inner coefficients. Put

$$a_i = \sqrt{A_i}(0, \ldots, 0, 1, -\alpha_{i,i+1}, \ldots, -\alpha_{in}). \tag{3.1}$$

Then $a_1, \dots, a_n \in \mathbb{R}^n$ are row vectors satisfying (**S'**) and (**N'**), and such that $Q = \sum_{i=1}^n a_i^t a_i$. Let

$$B^i = a_i^t a_i = A_i \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & 1 & -\alpha_{i,i+1} & \cdots & -\alpha_{in} \\ \mathbf{0} & -\alpha_{i,i+1} & \alpha_{i,i+1}\alpha_{i,i+1} & \cdots & \alpha_{i,i+1}\alpha_{in} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & -\alpha_{in} & \alpha_{in}\alpha_{i,i+1} & \cdots & \alpha_{in}\alpha_{in} \end{bmatrix}. \tag{3.2}$$

(Here the **0**'s are zero matrices and vectors of appropriate sizes.) Then $(B^1, \dots, B^n)$ satisfies (**r**), (**p**), (**s**), and (**n**).

For sufficiency, let $B^1, \dots, B^n$ be such that $Q = B^1 + \cdots + B^n$ and such that (**r**), (**p**), (**s**), and (**n**) hold. As B' has rank 1, we may write $B^i = a_i^t a_i$, where $a_{ii} \geq 0$. Then a_i satisfies (**N'**). To see that a_i satisfies (**S'**), let e_i be the ith unit vector in $\mathbb{R}^n$. Note that $e_i \in \operatorname{cone} D_i$ and that hence

$$e_i d^t \in \operatorname{cone}\{d_1 d_2^t \mid d_1, d_2 \in D_i\} \tag{3.3}$$

for any $d \in D_i$. From the fact that B^i satisfies (**s**) it follows that $(a_i^t a_i) \cdot D \geq 0$ for all $D \in \operatorname{cone}\{d_1 d_2^t \mid d_1, d_2 \in D_i\}$, and in particular that $(a_i e_i)(a_i d) \geq 0$ for all $d \in D_i$. Thus $a_i d \geq 0$ for all $d \in D_i$. □

So, for $(c_1, \dots, c_n) \in \mathbb{R}^n$, the minimum

$$\min\left\{ \sum_{i=1}^n c_i A_i \mid \sum_i A_i (x_i - \sum_{j>i} \alpha_{ij} x_j)^2 \text{ is KZ-reduced for some } \alpha_{ij}, A_n = 1 \right\} \tag{3.4}$$

equals

$$\min\left\{ \sum_k c_k B_{kk}^k \mid (B^1, \dots, B^n) \text{ satisfies } (\mathbf{r}), (\mathbf{p}), (\mathbf{s}), (\mathbf{n}), \text{ and } B_{nn}^n = 1 \right\}. \tag{3.5}$$

Here the extra condition at the end is added to remove scale invariance from the problem. Dropping the rank-1 constraint (**r**) yields a lower bound that is a semidefinite optimization problem:

$$z(c) := \min\left\{ \sum_{k=1}^n c_k B_{kk}^k \mid (B^1, \dots, B^n) \text{ satisfies } (\mathbf{p}), (\mathbf{s}), (\mathbf{n}), \text{ and } B_{nn}^n = 1 \right\}. \tag{3.6}$$

Note that it is possible to determine the value of (3.6) without knowing the Novikova sets X_i^* in advance, by using a cutting plane algorithm as follows. Replace in (3.6) the constraints (**n**) by the following for certain small sets X_i (for example, take $X_2 = X_2^*$ and the other X_i empty):

$$\sum_{i=k}^n \widetilde{x}^t B^i \widetilde{x} \geq B_{kk}^k \text{ for all } x \in X_{n-k+1}, \text{ for } k = 1, \dots, n-1. \tag{n'}$$

Repeatedly refine these constraints by solving the relaxation and finding, for some k, a nonzero vector $x \in \mathbb{Z}^{n-k+1}$ with $\sum_{i=k}^n \widetilde{x}^t B^i \widetilde{x} < B_{kk}^k$ for the optimal solution to the relaxation. Add this x to X^{n-k+1}. Eventually no such x will be found, and then the optimal solution to this relaxation will be equal to the optimal solution to (3.6).

One can use the techniques in the proof of Theorem 2.1 to bound the search space for these vectors.

A cutting plane algorithm may even be the only practical way to solve the relaxation for $n > 5$, since the cardinality of X_n^* seems to increase very rapidly with n. The following theorem, similar to Theorem 2.1, implies that such a cutting plane algorithm will finish.

THEOREM 3.2. *Let* $(B^1, \dots, B^n)$ *satisfy* (**p**), (**s**), *and suppose that*

$$\sum_{i=1}^{n} e_2^t B^i e_2 \geq B^1_{11}, \tag{3.7}$$

$$\sum_{i=k}^{n} \widetilde{x}^t B^i \widetilde{x} \geq B^k_{kk} \text{ for all nonzero } x \in \mathbb{Z}^{n-k+1}, k = 2, \dots, n-1. \tag{3.8}$$

Then there are only finitely many $x \in \mathbb{Z}^n \setminus \{0\}$ *such that* $\sum_{i=1}^n x^t B^i x < B^1_{11}$.

Compared to the method of Lasserre [9], in particular to a second-order moment relaxation of our polynomial optimization problem, our relaxation contains variables B^k_{ij} corresponding to products $a_{ki}a_{kj}$ but no variables corresponding to products $a_{ki}a_{lj}$ when $k \neq l$. Accordingly, we do not take products of linear inequalities $a_k d_1 \geq 0, a_l d_2 \geq 0$ into account.

4. Branch and bound. In this section we give an overview of the branching process. We refer to [17] for further details and a full implementation. In the definition of KZ-reducedness, the size-reduction requirement (**S**) asks that for $i = 1, \dots, n-1$ we have

$$(\alpha_{i,i+1}, \dots, \alpha_{in}) \in S^i := \left[0, \frac{1}{2}\right] \times \left[-\frac{1}{2}, \frac{1}{2}\right]^{n-i-1}. \tag{4.1}$$

There is nothing particular about the polyhedra S^i that makes the semidefinite relaxation (3.6) possible. Taking any set of polyhedra P^i instead of the S^i, a semidefinite lower bound $z(c, P^1, \dots, P^{n-1})$ analogous to (3.6) for

$$\min \Bigg\{ \sum c_i A_i \mid \sum_i A_i (x_i - \sum_{j>i} \alpha_{ij} x_j)^2 \text{ satisfies (**N**)}, \\ (\alpha_{i,i+1}, \dots, \alpha_{in}) \in P^i \text{ for } i = 1, \dots, n-1, \text{ and } A_n = 1 \Bigg\} \tag{4.2}$$

may be constructed. This new relaxation differs from (3.6) in the constraints (**s**). If the diameter of these polyhedra P^i is small, then the matrix B^i is close to a rank-1 matrix in the following sense. Suppose the width of P^i is small, i.e., for all j,

$$\max\{\alpha_{ij} \mid (\alpha_{i,i+1}, \dots, \alpha_{in}) \in P^i\} - \min\{\alpha_{ij} \mid (\alpha_{i,i+1}, \dots, \alpha_{in}) \in P^i\} < \varepsilon, \tag{4.3}$$

where we assume $\varepsilon < 1$ and $\max\{\alpha_{ij} \mid (\alpha_{i,i+1}, \dots, \alpha_{in}) \in P^i\} \leq 1/2$. Let $(B^1, \dots, B^n)$ be any feasible solution corresponding to $z(c, P^1, \dots, P^{n-1})$. Let $(\widetilde{B}^1, \dots, \widetilde{B}^n)$ be any feasible solution corresponding to $z(c, P^1, \dots, P^{n-1})$ such that $\widetilde{B}^i$ has rank 1. Then for all $j, k \in \{i, \dots, n\}$,

$$|B^i_{jk}/B^i_{ii} - \widetilde{B}^i_{jk}/\widetilde{B}^i_{ii}| \leq 2\varepsilon. \tag{4.4}$$

If we have a set of $(n-1)$-tuples of polyhedra $N = \{(P_s^1, \dots, P_s^{n-1}) \mid s = 1, \dots, t\}$ so that

$$S^1 \times \cdots \times S^{n-1} = \bigcup_{(P^1,\dots,P^{n-1})\in N} P^1 \times \cdots \times P^{n-1}, \tag{4.5}$$

then

$$\min\{z(c, P^1, \dots, P^{n-1}) \mid (P^1, \dots, P^{n-1}) \in N\} \tag{4.6}$$

is again a lower bound for (3.4). If we partition $S^1 \times \cdots \times S^{n-1}$ so that in each part the diameter of each of the P^i is small, then we would obtain a good lower bound. However, this would make the cardinality of N very large, even for moderately small ε. Therefore, we take an iterative approach. Initially we choose $N = \{(S^1, \dots, S^{n-1})\}$. Then we repeat the following. Suppose that the minimum of (4.6) is attained at $(P^1, \dots, P^{n-1}) \in N$. Then we choose an $i \in \{1, \dots, n-1\}$ and replace $(P^1, \dots, P^{n-1})$ in N by the two tuples

$$(P^1, \dots, P^{i-1}, Q, P^{i+1}, \dots, P^{n-1}) \text{ and } (P^1, \dots, P^{i-1}, Q', P^{i+1}, \dots, P^{n-1}), \tag{4.7}$$

where Q, Q' are polyhedra such that $P^i = Q \cup Q'$—so N retains property (4.5). This process of refining N continues until (4.6) is sufficiently close to the desired value or some other stopping criterion applies.

We choose i, Q, Q' with the aim of reducing the "distance" of an optimal solution to a rank-1 solution, as follows. If this optimal solution of the problem with optimum $z(c, P^1, \dots, P^{n-1})$ is $(B^1, \dots, B^n)$, then we take i, j so that

$$\sum_{k=i}^{n} \frac{1}{B_{ii}^i}(B_{ii}^i B_{jk}^i - B_{ij}^i B_{ik}^i) \tag{4.8}$$

is maximal. Then we put

$$\begin{aligned} Q &= \{(\alpha_{i,i+1}, \dots, \alpha_{in}) \in P^i \mid \alpha_{ij} \le \beta\}, \\ Q' &= \{(\alpha_{i,i+1}, \dots, \alpha_{in}) \in P^i \mid \alpha_{ij} \ge \beta\}, \end{aligned} \tag{4.9}$$

where β is (a rational number with modest denominator near) $-B_{ij}^i/B_{ii}^i$.

We have tried other methods for picking i, Q, Q', but this turned out to work best in practice, in the sense that the cardinality of N required to obtain a certain bound was the smallest we could attain. Only by constructing N by hand did we achieve a smaller set for one problem.

5. Rounding to obtain exact bounds. Every feasible solution y to the dual of (3.6) gives a lower bound on $z(c)$ and hence on the optimal solution to (3.4). A dual solution is feasible if and only if a number of matrices, say, $M_1(y), \dots, M_k(y)$, is positive semidefinite. In fact, in our computations we work only with solutions y that are *strictly* positive definite. This simplifies the verification of feasibility, but the crucial advantage is that it helps to counter the imprecision inherent in the computation with limited-precision floating point numbers.

In the dual of (3.6) such solutions can be obtained by replacing a dual constraint $M_i(y) \succeq 0$ with $M_i(y) \succeq \varepsilon I$, where I is an identity matrix of suitable dimension and ε is a small positive constant. Bringing this matrix to the other side, we get the perturbed constraint

$$M_i(y) - \varepsilon I \succeq 0, \tag{5.1}$$

which corresponds to a perturbation of the function that is being optimized in the primal problem. Again we refer the reader to [17] for further details.

A floating-point solution y to the perturbed problem can be approximated by a continued fraction expansion, a technique recently used in [15]. If this approximation $\widetilde{y}$ is sufficiently close to y, it might violate some of the perturbed dual constraints slightly, but it will be strictly feasible for the original problem. Positive definiteness can then be ascertained by evaluating $\sum_{i=1}^{k} \operatorname{rank}(M_i(\widetilde{y}))$ determinants.

Note that this approach can also be applied to find feasible solutions of the primal semidefinite problem but is quite useless when it comes to deriving an optimal solution of the original problem (3.4) or (4.2), that is, a solution that also satisfies the rank-1 constraints (**r**). This is of no concern when one is interested in lower bounds, but it is also interesting to find KZ-reduced forms that give a good upper bound. We do not have a very reliable automated method to obtain such forms—not even from the optimal solution of our branch and bound procedure, which is nonetheless close to rank 1 in the sense that (4.8) is small for all i, j.

6. New linear inequalities on the outer coefficients of KZ-reduced quadratic forms. We define

$$K_n := \operatorname{cone}\{A(q) \mid q \text{ is an } n\text{-ary KZ-reduced form}\}. \tag{6.1}$$

We have

$$K_n = \{x \in \mathbb{R}^n \mid (0, x) \in K_{n+1}\} \tag{6.2}$$

and

$$K_n = \{x \in \mathbb{R}^n \mid (x, y) \in K_{n+1} \text{ for some } y \in \mathbb{R}\}. \tag{6.3}$$

Table 6.1 gives several KZ-reduced forms, some of which come from [13], whereas others were found by manually rounding and tweaking primal solutions to (4.2) for suitably chosen c and polyhedra P^i. The format is as follows: the columns labeled "Outer" and "Inner" hold the vector and matrix

$$\begin{bmatrix} A_1 \\ \vdots \\ A_n \end{bmatrix}, \begin{bmatrix} 1 & -\alpha_{12} & \cdots & -\alpha_{1n} \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & -\alpha_{n-1,n} \\ 0 & \cdots & 0 & 1 \end{bmatrix}, \tag{6.4}$$

respectively.

By the first KZ-inequality, K_2 is contained in

$$K_2' := \left\{ (A_1, A_2) \in \mathbb{R}_+^2 \mid \ A_2 \geq \frac{3}{4} A_1 \right\}. \tag{6.5}$$

It follows from Table 6.1 that K_2 contains

TABLE 6.1
Some KZ-reduced forms.

Name	Outer	Inner	Form
E1	$\begin{bmatrix}1\end{bmatrix}$	$\begin{bmatrix}1\end{bmatrix}$	$\begin{bmatrix}1\end{bmatrix}$
E2	$\begin{bmatrix}1\\3/4\end{bmatrix}$	$\begin{bmatrix}1&-1/2\\0&1\end{bmatrix}$	$\frac{1}{2}\begin{bmatrix}2&1\\1&2\end{bmatrix}$
E3a	$\begin{bmatrix}1\\3/4\\2/3\end{bmatrix}$	$\begin{bmatrix}1&-1/2&1/2\\0&1&-1/3\\0&0&1\end{bmatrix}$	$\frac{1}{2}\begin{bmatrix}2&-1&1\\-1&2&-1\\1&-1&2\end{bmatrix}$
E3b	$\begin{bmatrix}1\\8/9\\2/3\end{bmatrix}$	$\begin{bmatrix}1&-1/3&-1/3\\0&1&-1/2\\0&0&1\end{bmatrix}$	$\frac{1}{3}\begin{bmatrix}3&-1&-1\\-1&3&-1\\-1&-1&3\end{bmatrix}$
E4a	$\begin{bmatrix}1\\3/4\\2/3\\1/2\end{bmatrix}$	$\begin{bmatrix}1&-1/2&1/2&1/2\\0&1&-1/3&-1/3\\0&0&1&-1/2\\0&0&0&1\end{bmatrix}$	$\frac{1}{2}\begin{bmatrix}2&-1&1&1\\-1&2&-1&-1\\1&-1&2&0\\1&-1&0&2\end{bmatrix}$
E4b	$\begin{bmatrix}1\\8/9\\2/3\\5/8\end{bmatrix}$	$\begin{bmatrix}1&-1/3&-1/3&1/3\\0&1&-1/2&1/2\\0&0&1&-1/4\\0&0&0&1\end{bmatrix}$	$\frac{1}{6}\begin{bmatrix}6&-2&-2&2\\-2&6&-2&2\\-2&-2&6&-3\\2&2&-3&6\end{bmatrix}$
E4c	$\begin{bmatrix}1\\15/16\\45/64\\5/8\end{bmatrix}$	$\begin{bmatrix}1&-1/4&-1/4&-1/4\\0&1&-1/2&-1/2\\0&0&1&-1/3\\0&0&0&1\end{bmatrix}$	$\frac{1}{32}\begin{bmatrix}32&-8&-8&-8\\-8&32&-13&-13\\-8&-13&32&2\\-8&-13&2&32\end{bmatrix}$
E5a	$\begin{bmatrix}1\\3/4\\2/3\\1/2\\1/2\end{bmatrix}$	$\begin{bmatrix}1&-1/2&1/2&1/2&1/2\\0&1&-1/3&-1/3&-1/3\\0&0&1&-1/2&1/4\\0&0&0&1&-1/2\\0&0&0&0&1\end{bmatrix}$	$\frac{1}{2}\begin{bmatrix}2&-1&1&1&1\\-1&2&-1&-1&-1\\1&-1&2&0&1\\1&-1&0&2&0\\1&-1&1&0&2\end{bmatrix}$
E5b	$\begin{bmatrix}1\\8/9\\2/3\\5/8\\15/32\end{bmatrix}$	$\begin{bmatrix}1&-1/3&-1/3&-1/3&-1/3\\0&1&-1/2&7/16&-1/2\\0&0&1&-3/8&-1/4\\0&0&0&1&-1/2\\0&0&0&0&1\end{bmatrix}$	$\frac{1}{6}\begin{bmatrix}6&-2&-2&-2&-2\\-2&6&-2&3&-2\\-2&-2&6&-2&1\\-2&3&-2&6&-2\\-2&-2&1&-2&6\end{bmatrix}$
E5c	$\begin{bmatrix}1\\3/4\\2/3\\5/8\\15/32\end{bmatrix}$	$\begin{bmatrix}1&-1/2&1/2&-1/2&-1/2\\0&1&-1/3&1/3&1/3\\0&0&1&-1/4&-1/4\\0&0&0&1&-1/2\\0&0&0&0&1\end{bmatrix}$	$\frac{1}{16}\begin{bmatrix}16&-8&8&-8&-8\\-8&16&-8&8&8\\8&-8&16&-8&-8\\-8&8&-8&16&1\\-8&8&-8&1&16\end{bmatrix}$
E5d	$\begin{bmatrix}1\\3/4\\3/4\\9/16\\1/2\end{bmatrix}$	$\begin{bmatrix}1&-1/2&1/4&-1/4&1/2\\0&1&-1/2&-1/2&0\\0&0&1&-1/2&1/2\\0&0&0&1&-1/3\\0&0&0&0&1\end{bmatrix}$	$\frac{1}{4}\begin{bmatrix}4&-2&1&-1&2\\-2&4&-2&-1&-1\\2&-2&4&-1&2\\-1&-1&-1&4&-2\\2&-1&2&-2&4\end{bmatrix}$

$$K_2'' := \operatorname{cone}\left\{\begin{bmatrix}0\\1\end{bmatrix}, \begin{bmatrix}1\\3/4\end{bmatrix}\right\}. \tag{6.6}$$

Since $K_2' = K_2''$, we have equality throughout in $K_2' \supseteq K_2 \supseteq K_2''$.

Also, K_3 is contained in

$$K_3' := \left\{(A_1, A_2, A_3) \in \mathbb{R}^3_+ \mid A_2 \geq \frac{3}{4}A_1,\ A_3 \geq \frac{3}{4}A_2,\ A_3 \geq \frac{2}{3}A_1\right\} \tag{6.7}$$

by the first and second KZ-inequalities, and K_3 contains

$$K_3'' := \operatorname{cone}\left\{\begin{bmatrix}0\\0\\1\end{bmatrix}, \begin{bmatrix}0\\1\\3/4\end{bmatrix}, \begin{bmatrix}1\\3/4\\2/3\end{bmatrix}, \begin{bmatrix}1\\8/9\\2/3\end{bmatrix}\right\}. \tag{6.8}$$

Again we have equality throughout in $K_3' \supseteq K_3 \supseteq K_3''$, as $K_3' = K_3''$.

For $n = 4$ the classical KZ-inequalities no longer suffice to determine K_n. By the first and second KZ-inequalities, K_4 is contained in

$$\left\{(A_1, A_2, A_3, A_4) \in \mathbb{R}_+^4 \mid A_{i+1} \geq \frac{3}{4}A_i,\ A_{i+2} \geq \frac{2}{3}A_i\right\}. \tag{6.9}$$

But the extremal vector $(1, \frac{8}{9}, \frac{2}{3}, \frac{16}{27})$ of this cone cannot be realized as the sequence of outer coefficients of a KZ-reduced form.

THEOREM 6.1. *Let $A_1, \ldots, A_4$ be the outer coefficients of a KZ-reduced form in four variables. Then*

$$-25A_1 - 36A_2 + 48A_3 + 40A_4 \geq -7 \cdot 10^{-6} A_4. \tag{6.10}$$

This theorem was proven by the branch-and-bound and rounding processes described in the previous sections. The data required to verify this theorem can be found in [17].

Thus $K_4 \subseteq K_4'$, where

$$K_4' := \left\{(A_1, A_2, A_3, A_4) \in \mathbb{R}_+^4 \mid\ A_{i+1} \geq \frac{3}{4}A_i,\ A_{i+2} \geq \frac{2}{3}A_i,\ (6.10)\right\}. \tag{6.11}$$

We conjecture that in the above theorem we even have

$$-25A_1 - 36A_2 + 48A_3 + 40A_4 \geq 0. \tag{6.12}$$

By Table 6.1, K_4 contains the cone

$$K_4'' := \operatorname{cone}\left\{\begin{bmatrix}0\\0\\0\\1\end{bmatrix}, \begin{bmatrix}0\\0\\1\\3/4\end{bmatrix}, \begin{bmatrix}0\\1\\3/4\\2/3\end{bmatrix}, \begin{bmatrix}0\\1\\8/9\\2/3\end{bmatrix}, \begin{bmatrix}1\\3/4\\2/3\\1/2\end{bmatrix}, \begin{bmatrix}1\\8/9\\2/3\\5/8\end{bmatrix}, \begin{bmatrix}1\\15/16\\45/64\\5/8\end{bmatrix}\right\}, \tag{6.13}$$

and we have

$$K_4'' = \left\{(A_1, A_2, A_3, A_4) \in \mathbb{R}_+^4 \mid\ A_{i+1} \geq \frac{3}{4}A_i,\ A_{i+2} \geq \frac{2}{3}A_i,\ (6.12)\right\}. \tag{6.14}$$

Hence K_4 is nearly determined by $K_4'' \supseteq K_4 \supseteq K_4'$, and our conjecture would imply $K_4 = K_4''$.

In dimension 5, we prove the following linear bounds.

THEOREM 6.2. *Let $A_1, \ldots, A_5$ be the outer coefficients of a KZ-reduced form in five variables. Then*

$$-5A_1 + 2A_4 + 8A_5 \geq -3 \cdot 10^{-4} A_5 \tag{6.15}$$

and

$$-4A_1 - 3A_3 + 4A_4 + 8A_5 \geq -5 \cdot 10^{-5} A_5. \tag{6.16}$$

TABLE 6.2
Incidences between some inequalities and elements of K_5. The rightmost column gives the dimension of the face of K_5 defined by the inequality.

Inequality	"Tight" forms	Rank
$-3A_1 + 4A_2 \geq 0$	E1, E2, E3a, E3b	4
$-3A_2 + 4A_3 \geq 0$	E1, E2, E4a, E5b	4
$-3A_3 + 4A_4 \geq 0$	E1, E3a, E4b, E4c	4
$-3A_4 + 4A_5 \geq 0$	E2, E3b, E4a, E5b	4
$-2A_1 + 3A_3 \geq 0$	E1, E2, E5a, E5b	4
$-2A_2 + 3A_4 \geq 0$	E1, E4a, E4b, E5a	4
$-2A_3 + 3A_5 \geq 0$	E3a, E3b, E5b	≥ 3
$-25A_1 - 36A_2 + 48A_3 + 40A_4 \geq 0$	E1, E5a, E5b	≥ 3
$-25A_2 - 36A_3 + 48A_4 + 40A_5 \geq 0$	E4a, E4b, E4c	≥ 3
$-5A_1 + 2A_4 + 8A_5 \geq 0$	E5a, E5b, E5c	≥ 3
$-4A_1 - 3A_3 + 4A_4 + 8A_5 \geq 0$	E5a, E5d	≥ 2

Of course, we conjecture

$$-5A_1 + 2A_4 + 8A_5 \geq 0 \tag{6.17}$$

and

$$-4A_1 - 3A_3 + 4A_4 + 8A_5 \geq 0. \tag{6.18}$$

As before, these inequalities describe a superset K_5' of K_5, and the forms of Table 6.1 generate a subset K_5'' of K_5. But there is now a fundamental discrepancy between K_5' and K_5''. Table 6.2 lists the known and conjectured inequalities for K_5 and with each inequality gives the forms of Table 6.1 that satisfy these inequalities with equality. Experimentation suggests that both inclusions in $K_5'' \subseteq K_5 \subseteq K_5'$ are strict (even if we replace, in the definition of K_5', the inequalities proven in Theorem 6.2 by their conjectured counterparts).

As an example, the four forms E5a, E5b, E5c, and E5d satisfy the following inequality:

$$-8A_1 - 3A_3 + 4A_4 + 16A_5 \geq 0. \tag{6.19}$$

One could conjecture that this is a facet of K_5. This is false, however; it is violated by the KZ-reduced form

$$\begin{bmatrix} 134 & -54 & -40 & -54 & 54 \\ -54 & 134 & -40 & 67 & -67 \\ -40 & -40 & 134 & -40 & -27 \\ -54 & 67 & -40 & 134 & -67 \\ 54 & -67 & -27 & -67 & 134 \end{bmatrix}. \tag{6.20}$$

We could obtain several other extreme forms in five variables and more valid inequalities, but we never reached a close approximation of K_5. Therefore, we publish only the two inequalities that seemed most relevant to the applications here. We maintain a list of certified inequalities at our website,[2] where our software [17] can also be found.

[2] http://www.win.tue.nl/kz/

TABLE 6.3

Relation between Hermite's constant and the approximation found. The first row gives the exact value of γ_n^n for $n \leq 8$, and the best known lower bound for $n = 9$. The second row gives the upper bound found using our approximation of K_n.

Dimension	1	2	3	4	5	6	7	8	9
(Lower bound on) γ_n^n	1	4/3	2	4	8	64/3	64	256	512
Upper bound	1	4/3	2	4	8.00005	21.3336	64.0012	256.008	1024.11

Even though we do not have a close approximation of K_5, we do have enough inequalities on the outer coefficients to bound Hermite's constant for $n \leq 8$ very well. Assuming the conjectured inequalities (6.12), (6.17), and (6.18), the upper bound on γ_n^n that would follow from the corresponding strengthening of (1.6) is exact for $n \leq 8$. Table 6.3 gives for $n = 1, \ldots, 8$ the known values of γ_n^n, and the upper bound on γ_n^n that follows from the proven inequalities (6.10), (6.15), and (6.16). In dimension 9 there is suddenly a huge gap between our upper bound and the best known lower bound. This gap is also larger than the gap obtained by the Cohn–Elkies bound. One or more new inequalities are needed to close this gap.

Blichfeldt observed in [1] that a tight upper bound on γ_n would follow for $n = 6, 7, 8$ from the two KZ-inequalities and "a certain inequality that we would reasonably expect to be true, namely, $A_{i+4} \geq \frac{1}{2}A_i$." But he immediately exhibits a set of forms showing that this inequality is false (the forms E5b and E5c of Table 6.1 are also counterexamples). Note that the inequalities we conjecture/approximate come near to this key inequality Blichfeldt suggests: (6.18) would imply that if $A_4 = \frac{3}{4}A_3$, then $A_5 \geq \frac{1}{2}A_1$, and (6.17) would imply that if $A_5 \leq (\frac{1}{2} - \epsilon)A_1$, then $A_4 \geq (\frac{1}{2} + 4\epsilon)A_1$.

7. The quality of block KZ-reduced lattice bases. If $L \subseteq \mathbb{R}^n$ is a full-dimensional lattice and $b_1, \ldots, b_n \in L$ are linearly independent vectors such that

$$L = \{x_1 b_1 + \cdots + x_n b_n \mid x_1, \ldots, x_n \in \mathbb{Z}\}, \tag{7.1}$$

then $b_1, \ldots, b_n$ is a *basis* of L. A basis of a lattice determines a positive definite quadratic form

$$q(x_1, \ldots, x_n) := \|x_1 b_1 + \cdots + x_n b_n\|^2. \tag{7.2}$$

A lattice basis $b_1, \ldots, b_n$ is said to be *KZ-reduced* if the associated form (7.2) is KZ-reduced.

Let $b_1^*, \ldots, b_n^*$ be the Gram–Schmidt orthogonalization of $b_1, \ldots, b_n$; that is, let $b_1^*, \ldots, b_n^*$ be pairwise orthogonal vectors so that

$$b_k = b_k^* - \sum_{i=1}^{k-1} \alpha_{ik} b_i^* \text{ for } k = 1, \ldots, n, \tag{7.3}$$

for some α_{ij}. Then these α_{ij} are exactly the inner coefficients of the associated form (7.2); and the outer coefficients of (7.2) satisfy

$$A_k = \|b_k^*\|^2. \tag{7.4}$$

So the classical KZ-inequalities and Theorems 6.1 and 6.2 can be read as inequalities relating the $\|b_i^*\|^2$ of a KZ-reduced lattice basis.

Block KZ-reduced lattice bases were introduced in [14] as a generalization of Lenstra–Lenstra–Lovasz (LLL-) reduced lattice bases. Such a basis gives a better estimate of the length of the shortest lattice vector and can still be computed in polynomial time when k is fixed. We say that a form

$$q(x_1,\dots,x_n)=\sum_{i=1}^{n} A_i\Big(x_i-\sum_{j>i}\alpha_{ij}x_j\Big)^2, \tag{7.5}$$

is *k-block KZ-reduced (k-BKZ-reduced)* if the derived forms

$$q_m^{m+k-1}(x_m,\dots,x_{m+k-1}):=\sum_{i=k}^{k+m-1} A_i\Big(x_i-\sum_{j=i+1}^{k+m-1}\alpha_{ij}x_j\Big)^2 \tag{7.6}$$

are KZ-reduced for $m=1,\dots,n-k+1$. Then a lattice basis is k-BKZ-reduced if the associated form is. In the remainder of this paper we will give some improved bounds on constants used in [14] for the analysis of the quality of k-BKZ-reduced lattice bases.

Let

$$\beta_{k,n}:=\max\frac{\|b_1^*\|^2}{\|b_n^*\|^2}, \tag{7.7}$$

where the maximum ranges over all k-BKZ-reduced lattice bases. Many of the useful properties of k-BKZ-reduced lattice bases are derived through upper bounds on $\beta_{k,n}$. As k increases toward n, $\beta_{k,n}$ is expected to decrease. Schnorr [14] defines $\alpha_k:=\beta_{k,k}$ and shows that

$$\beta_{k,1+m(k-1)}\le\alpha_k^m. \tag{7.8}$$

In terms of quadratic forms, one has

$$\beta_{k,n}=\max\left\{\frac{A_1}{A_n}\mid (A_1,\dots,A_n)=A(q),\ q \text{ a } k\text{-BKZ-reduced form}\right\} \tag{7.9}$$

and

$$\alpha_k=\max\left\{\frac{A_1}{A_k}\mid (A_1,\dots,A_k)=A(q),\ q \text{ a KZ-reduced form}\right\}. \tag{7.10}$$

It is immediate from the first KZ-inequality that $\alpha_2=\frac{4}{3}$ and from the second KZ-inequality that $\alpha_3=\frac{3}{2}$. A nonnegative combination of the inequalities (6.15) and $-\frac{3}{4}A_4+A_5\ge 0$ (the first KZ-inequality) is

$$-15A_1+32A_5\ge -9\cdot 10^{-4}A_5, \tag{7.11}$$

which implies

$$\alpha_5\le\frac{32}{15}+6\cdot 10^{-5}. \tag{7.12}$$

Since there exist KZ-reduced forms with $A_1/A_5=32/15$, we also have $\alpha_5\ge\frac{32}{15}$. For $k=4,5$, the bounds on $\beta_{k,n}$ that follow from (7.8) are only slightly weaker than those that follow directly from Theorems 6.1 and 6.2 by linear programming.

The limit

$$\widetilde{\beta}_k := \lim_{n\to\infty} \beta_{k,n}^{\frac{1}{n-1}} \tag{7.13}$$

also gives an indication of the relative effectiveness of k-BKZ-reduction. Observe that if an inequality $c_1 A_i + \cdots + c_k A_{i+k-1} \geq 0$ with $c_1 < 0$ holds for the outer coefficients of a KZ-reduced form in k variables, then $\widetilde{\beta}_k$ is bounded from above by the largest root of the polynomial $c_1 x^{k-1} + \cdots + c_k$. Thus the first KZ-inequality implies $\widetilde{\beta}_2 \leq 4/3 \approx 1.3333$, the second KZ-inequality implies $\widetilde{\beta}_3 \leq \sqrt{3/2} \approx 1.2247$, Theorem 6.1 implies $\widetilde{\beta}_4 \leq 1.2172$, and Theorem 6.2 (in particular (6.15)) implies $\widetilde{\beta}_5 \leq 1.2010$.

Acknowledgments. We thank Achill Schürmann and two anonymous referees for their comments, which improved the paper considerably.

REFERENCES

[1] H. F. Blichfeldt, *The minimum values of positive quadratic forms in six, seven, and eight variables*, Math. Z., 39 (1935), pp. 1–15.
[2] J. W. S. Cassels, *An Introduction to the Geometry of Numbers*, Classics in Mathematics, Springer-Verlag, Berlin, 1997. Corrected reprint of the 1971 edition.
[3] H. Cohn and N. Elkies, *New upper bounds on sphere packings* I, Ann. of Math. (2), 157 (2003), pp. 689–714.
[4] J. H. Conway and N. J. A. Sloane, *Sphere Packings, Lattices, and Groups*, 3rd ed., Grundlehren Math. Wiss. 290, Springer-Verlag, New York, 1999.
[5] B. N. Delone, *The St. Petersburg School of Number Theory*, Hist. Math. 26, AMS, Providence, RI, 2005.
[6] A. Korkine and G. Zolotareff, *Sur les formes quadratiques positives quaternaires*, Math. Ann., 5 (1872), pp. 581–583.
[7] A. Korkine and G. Zolotareff, *Sur les formes quadratiques*, Math. Ann., 6 (1873), pp. 366–389.
[8] A. Korkine and G. Zolotareff, *Sur les formes quadratiques positives*, Math. Ann., 11 (1877), pp. 242–292.
[9] J. B. Lasserre, *Global optimization with polynomials and the problem of moments*, SIAM J. Optim., 11 (2001), pp. 796–817.
[10] J. Martinet, *Perfect Lattices in Euclidean Spaces*, Grundlehren Math. Wiss. 327, Springer-Verlag, Berlin, 2003.
[11] N. V. Novikova, *Domains of Korkin-Zolotarev reduction of positive quadratic forms in $n \leq 8$ variables and reduction algorithms for these domains*, Dokl. Akad. Nauk SSSR, 270 (1983), pp. 48–51 (in Russian). English translation in Soviet Math. Doklady, 27 (1983), pp. 557–560.
[12] P. A. Parrilo, *Semidefinite programming relaxations for semialgebraic problems*, Math. Program., 96 (2003), pp. 293–320.
[13] S. S. Ryškov and E. P. Baranovskiĭ, *Classical methods of the theory of lattice packings*, Uspekhi Mat. Nauk, 34 (1979), pp. 3–63, 256 (in Russian). English translation in Russian Math. Surveys, 34 (1979), pp. 1–68.
[14] C.-P. Schnorr, *A hierarchy of polynomial time lattice basis reduction algorithms*, Theoret. Comput. Sci., 53 (1987), pp. 201–224.
[15] A. Schürmann and F. Vallentin, *Computational approaches to lattice packing and covering problems*, Discrete Comput. Geom., 35 (2006), pp. 73–116.
[16] S. H. M. van Zwam, *Properties of Lattices, a Semidefinite Programming Approach*, Master's thesis, Eindhoven University of Technology, Eindhoven, The Netherlands, 2005.
[17] S. H. M. van Zwam, *New Korkin–Zolotarev Inequalities: Implementation and Numerical Data*, SPOR report 2006-05, Eindhoven University of Technology, Eindhoven, The Netherlands, 2006. Available online at http://www.win.tue.nl/bs/spor/.

SIAM J. OPTIM.
Vol. 18, No. 2, pp. 379–388

CONVERGENCE OF THE GRADIENT SAMPLING ALGORITHM FOR NONSMOOTH NONCONVEX OPTIMIZATION*

KRZYSZTOF C. KIWIEL†

Abstract. We study the gradient sampling algorithm of Burke, Lewis, and Overton for minimizing a locally Lipschitz function f on $\mathbb{R}^n$ that is continuously differentiable on an open dense subset. We strengthen the existing convergence results for this algorithm and introduce a slightly revised version for which stronger results are established without requiring compactness of the level sets of f. In particular, we show that with probability 1 the revised algorithm either drives the f-values to $-\infty$, or each of its cluster points is Clarke stationary for f. We also consider a simplified variant in which the differentiability check is skipped and the user can control the number of f-evaluations per iteration.

Key words. generalized gradient, nonsmooth optimization, subgradient, gradient sampling, nonconvex

AMS subject classifications. 65K10, 90C26

DOI. 10.1137/050639673

1. Introduction.

In two recent papers [BLO02b, BLO05], Burke, Lewis, and Overton introduced and established convergence of the *gradient sampling* (GS) algorithm for minimizing a locally Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$ which is continuously differentiable on an open dense subset D of $\mathbb{R}^n$ and has bounded level sets.

At each iteration, the GS algorithm computes the gradient of f at the current iterate and at $m \geq n+1$ randomly generated nearby points. This bundle of gradients is used to find an approximate ϵ-steepest descent direction, where ϵ is the sampling radius, as the solution of a quadratic program. A standard Armijo line search along this direction produces a candidate for the next iterate, which is obtained by perturbing the candidate, if necessary, to stay in the set D where f is differentiable; the perturbation is random and small enough to maintain the Armijo sufficient descent property. The sampling radius ϵ may be fixed for all iterations or may be reduced dynamically. For ϵ fixed, the main convergence result of [BLO05, Theorem 3.4] established that, with probability 1, the GS algorithm generates a sequence with a cluster point that is ϵ-stationary for f (as defined in section 2). For ϵ reduced dynamically, the result of [BLO05, Theorem 3.8] established that if the GS algorithm converges to a point, this limit point is stationary for f with probability 1.

The GS algorithm is not only very interesting in theory (especially due to its ingenious use of gradients instead of subgradients [BLO02a]) but also widely applicable and robust in practice [BHLO06, BLO04, BLO05, Lew05].

This paper provides stronger convergence results for the GS algorithm. For ϵ fixed, we show that with probability 1 *every* cluster point of the GS algorithm is ϵ-stationary for f (see Theorem 3.6). For ϵ reduced dynamically, we show that with probability 1 every cluster point of a well-defined subsequence is stationary for f (see Theorem 3.4), without assuming that the whole sequence converges. In both cases, we show that suitable stopping criteria ensure with probability 1 that the algorithm

*Received by the editors September 6, 2005; accepted for publication (in revised form) October 20, 2006; published electronically May 7, 2007.
http://www.siam.org/journals/siopt/18-2/63967.html
†Systems Research Institute, Polish Academy of Sciences, Newelska 6, 01–447 Warsaw, Poland (kiwiel@ibspan.waw.pl).

terminates with the required "optimality certificate" of [BLO05, p. 768]; this practical aspect was not analyzed in [BLO05, section 3].

We also introduce a slight revision of the GS algorithm, in which the perturbation of the Armijo candidate is controlled by the current step size (instead of ϵ as in the original method; see (2.6)). This tiny modification enables us to derive much stronger convergence results; in particular, we can dispense with the assumption of [BLO05] that f has compact level sets. For ϵ fixed, we show that with probability 1 either the algorithm drives the f-values to $-\infty$, or every cluster point of a well-defined subsequence of its iterates is ϵ-stationary for f (see Theorem 3.5). For ϵ reduced dynamically, we show that with probability 1 the algorithm either drives the f-values to $-\infty$, or *each* of its cluster points is stationary for f (see Theorem 3.3); in a sense, this is the best result one can hope for. If $\inf f > -\infty$, in both cases suitable stopping criteria ensure with probability 1 that the algorithm terminates with the required "optimality certificate."

Our further modifications of the GS algorithm are intended to improve its performance in practice. Since the GS algorithm employs search directions of unit 2-norm, the number of f-evaluations per Armijo's line search can grow to infinity as the algorithm converges. To mitigate this drawback, we consider using an "unscaled" search direction, i.e., the negative of the convex combination of the gradients in the bundle whose norm is minimized. (This direction was used in [BLO02b, section 3] for a different line search.) The third alternative is to scale the direction so that its length equals ϵ and the Armijo line search is made within the ball in which gradient sampling occurs.

Finally, we introduce a lower bound on step sizes tested by the Armijo search, accepting a null step size when this bound is reached. Here the idea is simple: when the search direction is good enough, a step size close to our lower bound should work, whereas if the search direction is poor, the Armijo search will produce a tiny step size anyway. In our limited Armijo line search (see Procedure 4.3), the number of f-evaluations can be controlled by the choice of an initial step size; in an extreme version, just one evaluation occurs. Further, for our limited line search there is no longer any need for keeping the iterates in the set D where f is differentiable. Skipping the differentiability check makes life easier for the user who provides gradient values and brings the simplified algorithm closer to the version implemented and tested in [BLO05, section 4].

Among other algorithms for minimizing locally Lipschitz functions, we should mention bundle methods (see the references in [BLO05, Kiw96]). Bundle methods require the computation of a single subgradient at each trial point in addition to the objective value. They generate search directions by solving quadratic programs based on accumulated subgradients and employ line searches which either produce descent or find a new subgradient that modifies the next search direction. At first sight, they have little in common with the GS algorithm, which does not accumulate gradients. We believe, however, that deeper understanding of their similarities and differences should lead to new variants. The first step in this direction is made here: the proof technique of section 3 is borrowed from [Kiw96, section 3], and the limited line search of section 4.3 is inspired by null steps of bundle methods. We defer consideration of gradient sampling in bundle methods, as well as numerical comparisons, to future work.

The paper is organized as follows. A slightly revised version of the GS algorithm is presented in section 2. A convergence analysis of the original and revised versions is given in section 3. Various modifications are discussed in section 4.

2. The GS algorithm. As in [BLO05], we assume that the objective function $f : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz continuous and continuously differentiable on an open dense subset D of $\mathbb{R}^n$. The Clarke *subdifferential* [Cla83] of f at any point x is given by

$$\bar{\partial} f(x) = \mathrm{co}\big\{ \lim_j \nabla f(y^j) : y^j \to x, y^j \in D \big\},$$

where co denotes the convex hull, and the Clarke ϵ-*subdifferential* [Gol77] by

$$\bar{\partial}_\epsilon f(x) := \mathrm{co}\, \bar{\partial} f(B(x, \epsilon)), \tag{2.1}$$

where $B(x, \epsilon) := \{y : |y - x| \le \epsilon\}$ is the ball centered at x with radius $\epsilon \ge 0$ and $|\cdot|$ is the 2-norm. The Clarke ϵ-subdifferential $\bar{\partial}_\epsilon f(x)$ is approximated by the set of [BLO05]

$$G_\epsilon(x) := \mathrm{cl}\,\mathrm{co}\, \nabla f(B(x, \epsilon) \cap D), \tag{2.2}$$

since $G_\epsilon(x) \subset \bar{\partial}_\epsilon f(x)$, and $\bar{\partial}_{\epsilon_1} f(x) \subset G_{\epsilon_2}(x)$ for $0 \le \epsilon_1 < \epsilon_2$. We say that a point x is *stationary* for f if $0 \in \bar{\partial} f(x)$; x is called ϵ-*stationary* for f if $0 \in \bar{\partial}_\epsilon f(x)$.

We now state a slightly revised version of the GS algorithm [BLO05, section 2]. In particular, we do not require that the starting point $x^1 \in D$ is such that the level set $\{x : f(x) \le f(x^1)\}$ is compact. For a closed convex set G, $\mathrm{Proj}(0 \,|\, G)$ is its minimum-norm element.

Algorithm 2.1 (revised GS algorithm).

Step 0 (initialization). Select an initial point $x^1 \in D$, optimality tolerances $\nu_{\mathrm{opt}}, \epsilon_{\mathrm{opt}} \ge 0$, line search parameters β, γ in $(0, 1)$, reduction factors μ, θ in $(0, 1]$, a sampling radius $\epsilon_1 > 0$, a stationarity target $\nu_1 \ge 0$, and a sample size $m \ge n + 1$. Set $k := 1$.

Step 1 (approximate the Clarke ϵ-subdifferential by gradient sampling). Let $\{x^{ki}\}_{i=1}^m$ be sampled independently and uniformly from $B(x^k, \epsilon_k)$. If $\{x^{ki}\}_{i=1}^m \not\subset D$, then stop; otherwise, set

$$G_k := \mathrm{co}\big\{\nabla f(x^k), \nabla f(x^{k1}), \ldots, \nabla f(x^{km})\big\}. \tag{2.3}$$

Step 2 (direction finding). Set $g^k := \mathrm{Proj}(0 \,|\, G_k)$.

Step 3 (stopping criterion). If $|g^k| \le \nu_{\mathrm{opt}}$ and $\epsilon_k \le \epsilon_{\mathrm{opt}}$, terminate.

Step 4 (sampling radius update). If $|g^k| \le \nu_k$, set $\nu_{k+1} := \theta\nu_k$, $\epsilon_{k+1} := \mu\epsilon_k$, $t_k := 0$, and $x^{k+1} := x^k$ and go to Step 7. Otherwise, set $\nu_{k+1} := \nu_k$, $\epsilon_{k+1} := \epsilon_k$, and

$$d^k := -g^k/|g^k|. \tag{2.4}$$

Step 5 (line search). Set the step size

$$t_k := \max\big\{ t : f(x^k + td^k) < f(x^k) - \beta t|g^k|, t \in \{1, \gamma, \gamma^2, \ldots\} \big\}. \tag{2.5}$$

Step 6 (updating). If $x^k + t_k d^k \in D$, set $x^{k+1} := x^k + t_k d^k$. Otherwise, let x^{k+1} be any point in D satisfying

$$f(x^{k+1}) < f(x^k) - \beta t_k |g^k|, \tag{2.6a}$$

$$|x^k + t_k d^k - x^{k+1}| \le \min\{t_k, \epsilon_k\}. \tag{2.6b}$$

Step 7. Increase k by 1 and go to Step 1.

The algorithm keeps every iterate x^k in the set D. At Step 2, g^k is characterized by $g^k \in G_k$ and $\langle g, g^k\rangle \geq |g^k|^2$ for all $g \in G_k$; since $\nabla f(x^k) \in G_k$ by (2.3), (2.4) yields $\langle \nabla f(x^k), d^k\rangle \leq -|g^k|$. Hence the Armijo line search (2.5) is well defined, because there is $\bar{t} > 0$ such that $f(x^k + td^k) < f(x^k) - \beta t|g^k|$ for all $t \in (0, \bar{t})$.

The only significant difference between Algorithm 2.1 and the original GS algorithm [BLO05, section 2] lies in the slightly stronger requirement (2.6). Namely, if $x^k + t_k d^k \notin D$, x^{k+1} can be found as follows. For $i = 1, 2, \ldots$, sample x^{k+1} from a uniform distribution on $B(x^k + t_k d^k, \min\{t_k, \epsilon_k\}/i)$ until $x^{k+1} \in D$ and (2.6a) holds. By (2.5) and the continuity of f, this procedure terminates with probability 1. In contrast, the original GS algorithm requires finding $\hat{x}^k$ in $B(x^k, \epsilon_k)$ such that $\hat{x}^k + t_k d^k \in D$ and (2.6a) holds for $x^{k+1} := \hat{x}^k + t_k d^k$; to this end, one can sample $\hat{x}^k$ from a uniform distribution on $B(x^k, \epsilon_k/i)$ until these requirements are met. Further, if (2.6) holds, then $\hat{x}^k := x^{k+1} - t_k d^k$ satisfies the requirements of the original GS algorithm. This is the only reason for including ϵ_k in (2.6b). On the other hand, the presence of t_k in (2.6b) yields $|x^{k+1} - x^k| \leq 2t_k$ (using $|d^k| = 1$ by (2.4)) and hence the highly useful consequence of (2.6a)

$$f(x^{k+1}) \leq f(x^k) - \beta\tfrac{1}{2}|x^{k+1} - x^k||g^k|. \tag{2.7}$$

Note that this key inequality (2.7) holds also when $x^{k+1} := x^k + t_k d^k$ at Step 6 (thanks to (2.5)) or when $x^{k+1} := x^k$ at Step 4.

The stopping criterion of Step 3 delivers the "optimality certificate" of [BLO05, p. 768]: the final values of $|g^k|$ and ϵ_k provide an estimate of nearness to Clarke stationarity.

3. Convergence analysis. We start with two technical lemmas. The first lemma on approximate least-norm elements is a simplified version of [BLO05, Lemma 3.1].

LEMMA 3.1. *Let $\emptyset \neq C \subset \mathbb{R}^n$ be compact convex and $\beta \in (0,1)$. If $0 \notin C$, there exists $\delta > 0$ such that $u, v \in C$ and $|u| \leq \operatorname{dist}(0 \,|\, C) + \delta$ imply $\langle v, u\rangle > \beta|u|^2$.*

Proof. If the assertion were false, we could pick two sequences $\{u^i\}, \{v^i\} \subset C$ satisfying $|u^i| \leq \operatorname{dist}(0 \,|\, C) + 1/i$ and $\langle v^i, u^i\rangle \leq \beta|u^i|^2$. By compactness, we may assume $u^i \to \bar{u} \in C$, $v^i \to \bar{v} \in C$; thus $\langle \bar{v}, \bar{u}\rangle \leq \beta|\bar{u}|^2$. However, $\bar{u} = \operatorname{Proj}(0 \,|\, C) \neq 0$ satisfies $\langle v, \bar{u}\rangle \geq |\bar{u}|^2$ for all $v \in C$, a contradiction. ∎

The next lemma recalls from [BLO05, Lemma 3.2] basic properties of the set of points close to a given point $\bar{x}$ that can be used to provide a δ-approximation to the least-norm element of $G_\epsilon(\bar{x})$; its second part summarizes some useful ideas from the proof of [BLO05, Theorem 3.4]. For $\epsilon, \delta > 0$ and $\bar{x}, x \in \mathbb{R}^n$, using the measure of proximity to ϵ-stationarity

$$\rho_\epsilon(\bar{x}) := \operatorname{dist}(0 \,|\, G_\epsilon(\bar{x})), \tag{3.1}$$

let

$$D_\epsilon^m(x) := \prod_1^m (B(x, \epsilon) \cap D) \subset \prod_1^m \mathbb{R}^n \tag{3.2}$$

and

$$V_\epsilon(\bar{x}, x, \delta) := \big\{ (y^1, \ldots, y^m) \in D_\epsilon^m(x) : \operatorname{dist}(0 \,|\, \operatorname{co}\{\nabla f(y^i)\}_{i=1}^m) \leq \rho_\epsilon(\bar{x}) + \delta \big\}. \tag{3.3}$$

LEMMA 3.2. *Let $\epsilon > 0$ and $\bar{x} \in \mathbb{R}^n$.*

(i) *For any $\delta > 0$, there is $\tau > 0$ and a nonempty open set $\bar{V}$ satisfying $\bar{V} \subset V_\epsilon(\bar{x}, x, \delta)$ for all $x \in B(\bar{x}, \tau)$, and* $\operatorname{dist}(0 \,|\, \operatorname{co}\{\nabla f(y^i)\}_{i=1}^m) \le \rho_\epsilon(\bar{x}) + \delta$ *for all* $(y^1, \ldots, y^m) \in \bar{V}$.

(ii) *Assuming $0 \notin G_\epsilon(\bar{x})$, pick $\delta > 0$ as in Lemma* 3.1 *for $C := G_\epsilon(\bar{x})$, and then τ and $\bar{V}$ as in statement* (i). *Suppose at iteration k of Algorithm* 2.1, *Step* 5 *is reached with $x^k \in B(\bar{x}, \min\{\tau, \epsilon/3\})$, $\epsilon_k = \epsilon$, and $(x^{k1}, \ldots, x^{km}) \in \bar{V}$. Then $t_k \ge \min\{1, \gamma\epsilon/3\}$.*

(iii) *If $\underline{\lim}_k \max\{|x^k - \bar{x}|, |g^k|, \epsilon_k\} = 0$ with $g^k \in \bar{\partial}_{\epsilon_k} f(x^k)$ for all k, then $0 \in \bar{\partial} f(\bar{x})$.*

Proof. (i) Let $u \in \operatorname{co} \nabla f(B(\bar{x}, \epsilon) \cap D)$ be such that $|u| < \rho_\epsilon(\bar{x}) + \delta$. Then Carathéodory's theorem [Roc70] implies the existence of $(\bar{x}^1, \ldots, \bar{x}^m) \in D_\epsilon^m(\bar{x})$ and $\bar{\lambda} \in \mathbb{R}_+^m$ with $\sum_{i=1}^m \bar{\lambda}_i = 1$ such that $u = \sum_{i=1}^m \bar{\lambda}_i \nabla f(\bar{x}^i)$. Since f is continuously differentiable on the open set D, there is $\bar{\epsilon} \in (0, \epsilon)$ such that the set $\bar{V} := \prod_{i=1}^m \operatorname{int} B(\bar{x}^i, \bar{\epsilon})$ lies in $D_{\epsilon-\bar{\epsilon}}^m(\bar{x})$ and $|\sum_{i=1}^m \bar{\lambda}_i \nabla f(y^i)| < \rho_\epsilon(\bar{x}) + \delta$ for all $(y^1, \ldots, y^m) \in \bar{V}$. Hence for all $x \in B(\bar{x}, \tau)$ with $\tau := \bar{\epsilon}$, the fact that $B(\bar{x}, \epsilon - \bar{\epsilon}) \subset B(x, \epsilon)$ yields $\bar{V} \subset V_\epsilon(\bar{x}, x, \delta)$ by the definitions (3.2)–(3.3).

(ii) Let $\hat{G}_k := \operatorname{co}\{\nabla f(x^{ki})\}_{i=1}^m$. Since $(x^{k1}, \ldots, x^{km}) \in \bar{V} \subset V_\epsilon(\bar{x}, \bar{x}, \delta)$ in statement (i), we get $\operatorname{dist}(0 \,|\, \hat{G}_k) \le \rho_\epsilon(\bar{x}) + \delta$ and $\hat{G}_k \subset G_\epsilon(\bar{x})$ from (3.3), (3.2), and (2.2). We also have $\nabla f(x^k) \in G_\epsilon(\bar{x})$ from $x^k \in B(\bar{x}, \epsilon/3) \cap D$. Thus, by (2.3) and the construction of g^k at Step 2, $g^k \in G_\epsilon(\bar{x})$ and $|g^k| \le \rho_\epsilon(\bar{x}) + \delta$. Hence by (3.1) and the choice of δ in Lemma 3.1,

$$\langle v, g^k \rangle > \beta |g^k|^2 \quad \text{for all } v \in G_\epsilon(\bar{x}). \tag{3.4}$$

Suppose for contradiction that $t_k < \min\{1, \gamma\epsilon/3\}$. Then by construction (cf. (2.5))

$$-\beta\gamma^{-1} t_k |g^k| \le f(x^k + \gamma^{-1} t_k d^k) - f(x^k),$$

whereas Lebourg's mean value theorem (cf. [Cla83, Theorem 2.3.7]) yields the existence of $\tilde{x}^k \in [x^k + \gamma^{-1} t_k d^k, x^k]$ and $v^k \in \bar{\partial} f(\tilde{x}^k)$ such that

$$f(x^k + \gamma^{-1} t_k d^k) - f(x^k) = \gamma^{-1} t_k \langle v^k, d^k \rangle.$$

Hence using $d^k := -g^k/|g^k|$ gives $\langle v^k, g^k \rangle \le \beta |g^k|^2$, and so $v^k \notin G_\epsilon(\bar{x})$ by (3.4). But $\gamma^{-1} t_k |d^k| < \epsilon/3$ and $|x^k - \bar{x}| \le \epsilon/3$ imply $\tilde{x}^k \in B(\bar{x}, 2\epsilon/3)$ and thus $v^k \in G_\epsilon(\bar{x})$, a contradiction.

(iii) Note that $g^k \in \bar{\partial}_{\epsilon_k} f(x^k)$ at Step 2 by (2.1), whereas $\bar{\partial}_\cdot f(\cdot)$ is closed. ∎

As discussed in section 2, Algorithm 2.1 is a special case of the GS algorithm, which in turn corresponds to removing t_k in the right-hand side of (2.6b) and requiring that the level set $\{x : f(x) \le f(x^1)\}$ be bounded. Therefore, we give convergence results separately for Algorithm 2.1 and the original GS algorithm. We start with the case where ϵ_k and ν_k are allowed to decrease.

THEOREM 3.3. *Let $\{x^k\}$ be a sequence generated by Algorithm* 2.1 *with $\nu_1 > \nu_{\text{opt}} = \epsilon_{\text{opt}} = 0$ and $\mu, \theta < 1$. With probability* 1 *the algorithm does not stop, and either $f(x^k) \downarrow -\infty$, or $\nu_k \downarrow 0$, $\epsilon_k \downarrow 0$ and every cluster point of $\{x^k\}$ is stationary for f.*

Proof. (i) Since termination in Step 1 has zero probability, we may assume it does not occur. Similarly, if $f(x^k) \downarrow -\infty$, there is nothing to prove, and so assume $\inf_k f(x^k) > -\infty$. Then summing $\beta t_k |g^k| \le f(x^k) - f(x^{k+1})$ (cf. (2.6a)) and relation (2.7) gives

$$\sum_{k=1}^{\infty} t_k |g^k| < \infty, \tag{3.5}$$

$$\sum_{k=1}^{\infty} |x^{k+1} - x^k||g^k| < \infty. \tag{3.6}$$

(ii) Suppose there is k_1, $\bar\nu > 0$, and $\bar\epsilon > 0$ such that $\nu_k = \bar\nu$ and $\epsilon_k = \bar\epsilon$ for all $k \ge k_1$. Using $|g^k| \ge \bar\nu$ in (3.5)–(3.6) yields $t_k \to 0$, $\sum_k |x^{k+1} - x^k| < \infty$, and hence the existence of a point $\bar x$ such that $x^k \to \bar x$. Let $\epsilon := \bar\epsilon$. First, suppose $0 \notin G_\epsilon(\bar x)$. For δ, τ, and $\bar V$ chosen as in Lemma 3.2(ii), we can pick $k_2 \ge k_1$ such that $x^k \in B(\bar x, \min\{\tau, \epsilon/3\})$ and $t_k < \min\{1, \gamma\epsilon/3\}$ yield $(x^{k1}, \dots, x^{km}) \notin \bar V$ for all $k \ge k_2$. This event has probability 0, since for each $k \ge k_2$, $(x^{k1}, \dots, x^{km})$ is sampled independently and uniformly from $D_\epsilon^m(x^k)$, which contains the open set $\bar V \ne \emptyset$. Second, suppose $0 \in G_\epsilon(\bar x)$. For $\delta := \bar\nu/2$ and $\tau, \bar V$ chosen as in Lemma 3.2(i), we can pick $k_3 \ge k_1$ such that $x^k \in B(\bar x, \tau)$, $\bar\nu \le |g^k| \le \operatorname{dist}(0 \,|\, \operatorname{co}\{\nabla f(x^{ki})\}_{i=1}^m)$, and $\rho_\epsilon(\bar x) = 0$ imply $(x^{k1}, \dots, x^{km}) \notin \bar V$ for all $k \ge k_3$. This event has probability 0 as well.

(iii) Consider the event where $\nu_k \downarrow 0$, $\epsilon_k \downarrow 0$, and $\{x^k\}$ has a cluster point $\bar x$. If $x^k \to \bar x$, $0 \in \bar\partial f(\bar x)$ by Lemma 3.2(iii). If $x^k \not\to \bar x$, we claim that $\underline{\lim}_k \max\{|x^k - \bar x|, |g^k|\} = 0$. Otherwise, there exist $\bar\nu > 0$, $\bar k$, and an infinite set $K := \{k : k \ge \bar k, |x^k - \bar x| \le \bar\nu\}$ such that $|g^k| > \bar\nu$ for all $k \in K$, and so (3.6) gives $\sum_{k\in K} |x^{k+1} - x^k| < \infty$. Since $x^k \not\to \bar x$, there is $\epsilon > 0$ such that for each $k \in K$ with $|x^k - \bar x| \le \bar\nu/2$ there exists $k' > k$ satisfying $|x^{k'} - x^k| > \epsilon$ and $|x^i - \bar x| \le \bar\nu$ for all $k \le i < k'$. Therefore, by the triangle inequality, we have $\epsilon < |x^{k'} - x^k| \le \sum_{i=k}^{k'-1} |x^{i+1} - x^i|$ with the right-hand side being less than ϵ for large $k \in K$ from $\sum_{k\in K} |x^{k+1} - x^k| < \infty$, a contradiction. Therefore, $\underline{\lim}_k \max\{|x^k - \bar x|, |g^k|\} = 0$ yields $0 \in \bar\partial f(\bar x)$ by Lemma 3.2(iii). □

THEOREM 3.4. *Let $\{x^k\}$ be a sequence generated by the original GS algorithm with $\nu_1 > \nu_{\rm opt} = \epsilon_{\rm opt} = 0$ and $\mu, \theta < 1$. Suppose the level set $\{x : f(x) \le f(x^1)\}$ is bounded. Then with probability 1 the algorithm does not stop, $\nu_k \downarrow 0$, $\epsilon_k \downarrow 0$, there is a subsequence $K \subset \{1, 2, \dots\}$ such that $g^k \xrightarrow{K} 0$, and every cluster point of $\{x^k\}_{k\in K}$ is stationary for f.*

Proof. It suffices to reconsider part (ii) of the proof of Theorem 3.3 (since, for $\nu_k \downarrow 0$, we can take $K := \{k : \nu_{k+1} < \nu_k\}$).

Thus suppose there is k_1, $\bar\nu > 0$, and $\bar\epsilon > 0$ such that $\nu_k = \bar\nu$ and $\epsilon_k = \bar\epsilon$ for all $k \ge k_1$. Using $|g^k| \ge \bar\nu$ in (3.5) yields $t_k \to 0$. Since $\{f(x^k)\}$ is decreasing and the set $\{x : f(x) \le f(x^1)\}$ is compact, there are a set $J \subset \{1, 2, \dots\}$ and a point $\bar x$ such that $x^k \xrightarrow{J} \bar x$. Since $t_k \xrightarrow{J} 0$ as well, arguing as in part (ii) of the proof of Theorem 3.3 we deduce the existence of k_4 and an open set $\bar V \ne \emptyset$ such that $(x^{k1}, \dots, x^{km}) \notin \bar V \subset D_\epsilon^m(x^k)$ for all $k \ge k_4$, $k \in J$, and again conclude that this event has probability 0. □

Our convergence results for fixed sampling radius follow.

THEOREM 3.5. *Let $\{x^k\}$ be a sequence generated by Algorithm 2.1 with $\nu_1 = \nu_{\rm opt} = 0$, $\epsilon_1 = \epsilon_{\rm opt} = \epsilon > 0$, and $\mu = 1$. With probability 1 either the algorithm terminates at some iteration k with $0 \in G_\epsilon(x^k)$, or $f(x^k) \downarrow -\infty$, or there is a subsequence $K \subset \{1, 2, \dots\}$ such that $g^k \xrightarrow{K} 0$ and every cluster point $\bar x$ of $\{x^k\}_{k\in K}$ satisfies $0 \in \bar\partial_\epsilon f(\bar x)$.*

Proof. If the algorithm terminates at iteration k, then with probability 1 it does so at Step 3 with $0 = g^k \in G_\epsilon(x^k)$. Hence we may assume that no termination occurs and $\inf_k f(x^k) > -\infty$.

By the proof of Theorem 3.3, the event $\bar\nu := \inf_k |g^k| > 0$ has probability 0. In the remaining case of $\inf_k |g^k| = 0$, the conclusion follows from the closedness of $\bar\partial_\epsilon f(\cdot)$. □

THEOREM 3.6. *Let $\{x^k\}$ be a sequence generated by the original GS algorithm with $\nu_1 = \nu_{\text{opt}} = 0$, $\epsilon_1 = \epsilon_{\text{opt}} = \epsilon > 0$, and $\mu = 1$. Suppose the set $\{x : f(x) \le f(x^1)\}$ is bounded. With probability 1 either the algorithm terminates at some iteration k with $0 \in G_\epsilon(x^k)$, or $g^k \to 0$ and every cluster point $\bar{x}$ of $\{x^k\}$ satisfies $0 \in \bar{\partial}_\epsilon f(\bar{x})$.*

Proof. Arguing by contradiction, it suffices to consider the case where there are a set $J \subset \{1, 2, \dots\}$ and $\bar{\nu} > 0$ such that $\inf_{k \in J} |g^k| \ge \bar{\nu}$. Since $\{f(x^k)\}$ is decreasing and the set $\{x : f(x) \le f(x^1)\}$ is compact, we may assume with no loss of generality that there is a point $\bar{x}$ such that $x^k \xrightarrow{J} \bar{x}$. Since (3.5) gives $t_k \xrightarrow{J} 0$, arguing as in part (ii) of the proof of Theorem 3.3 we deduce the existence of k_5 and an open set $\bar{V} \ne \emptyset$ such that $(x^{k1}, \dots, x^{km}) \notin \bar{V} \subset D_\epsilon^m(x^k)$ for all $k \ge k_5$, $k \in J$. This event has probability 0, since for each k, $(x^{k1}, \dots, x^{km})$ is sampled independently and uniformly from $D_\epsilon^m(x^k)$. ∎

A few comments and comparisons with the results of [BLO05, section 3] are in order.

Remark 3.7.

(i) Since the framework of [BLO05, section 3] requires compactness of the level set $\{x : f(x) \le f(x^1)\}$, it has no results comparable to our Theorems 3.3 and 3.5.

(ii) Theorem 3.3 is essentially the best one can hope for. In particular, it implies that for positive optimality tolerances ν_{opt} and ϵ_{opt}, with probability 1 either $f(x^k) \downarrow -\infty$, or the algorithm terminates with the required "optimality certificate" of [BLO05, p. 768].

(iii) Theorem 3.3 is stronger than Theorem 3.4. Of course, Theorem 3.3 relies on our inclusion of t_k in the right-hand side of (2.6b), but this should be cheap in practice. With this fairly mild qualification, Theorem 3.3 gives a positive answer to the final open question of [BLO05, section 3] on whether all cluster points of the algorithm are stationary.

(iv) Theorem 3.4 subsumes [BLO05, Theorem 3.8], which assumes that $\{x^k\}$ converges.

(v) Theorem 3.6 subsumes [BLO05, Theorem 3.4], which asserts only the existence of a subsequence $K \subset \{1, 2, \dots\}$ such that $\rho_\epsilon(x^k) \xrightarrow{K} 0$ and every cluster point $\bar{x}$ of $\{x^k\}_{k \in K}$ satisfies $0 \in \bar{\partial}_\epsilon f(\bar{x})$, without showing that $\inf_k |g^k| = 0$. In contrast, Theorem 3.6 implies that for a positive optimality tolerance ν_{opt}, with probability 1 the algorithm terminates when the required "optimality certificate" is reached (similarly for Theorem 3.5 if $\inf f > -\infty$). A result similar to Theorem 3.6 is given in [BLO05, part 1 of Corollary 3.5] only for the case where the objective f is continuously differentiable everywhere. Finally, Theorem 3.6 disproves the conjecture raised in the open question number 2 at the end of [BLO05, section 3] that a counterexample with $\overline{\lim}_{k \in K} |g^k| > 0$ should exist.

4. Modifications. Although our revision of Step 6 yields stronger theoretical results, it makes no difference to its implementation in practice when, as explained in [BLO05, section 4], it is not possible or practical to check whether the iterates lie in the set D where f is differentiable. Further, the implementation of [BLO05, section 4] obtained best results for the Armijo parameter $\beta = 0$ (although $\beta > 0$ is required in theory). Thus there is still the need for further study of line searches. In this section we propose several themes, supported by theory, that might prove useful in improving the practical performance of the method.

4.1. Nonnormalized search directions. Since the GS algorithm employs search directions $d^k := -g^k/|g^k|$ of unit norm, the number of f-evaluations per

Armijo's line search (cf. (2.5)) can grow to infinity. This will happen in the generic case where $x^{k+1} = x^k + t_k d^k$ for almost all k and $t_k = |x^{k+1} - x^k| \to 0$ (e.g., $\{x^k\}$ converges). To mitigate this drawback, let us consider using $d^k := -g^k$ as in the steepest descent method with $d^k = -\nabla f(x^k)$ in the smooth case.

Formally, suppose relations (2.4)–(2.6) in Algorithm 2.1 are replaced by

$$d^k := -g^k, \tag{4.1}$$

$$t_k := \max\{\, t : f(x^k + td^k) < f(x^k) - \beta t|g^k|^2, t \in \{1, \gamma, \gamma^2, \dots\}\,\}, \tag{4.2}$$

$$f(x^{k+1}) < f(x^k) - \beta t_k |g^k|^2, \tag{4.3a}$$

$$|x^k + t_k d^k - x^{k+1}| \le \min\{t_k, \epsilon_k\}|d^k|. \tag{4.3b}$$

Then (2.7) still holds, since $|x^{k+1} - x^k| \le 2t_k|d^k| = 2t_k|g^k|$. Lemma 3.2(ii) is replaced by the following.

LEMMA 4.1. *Let $\epsilon > 0$ and $\bar{x} \in \mathbb{R}^n$. Assuming $0 \notin G_\epsilon(\bar{x})$, pick $\delta > 0$ as in Lemma 3.1 for $C := G_\epsilon(\bar{x})$, and then τ and $\bar{V}$ as in Lemma 3.2(i). Suppose $x^k \in B(\bar{x}, \min\{\tau, \epsilon/3\})$, $\epsilon_k = \epsilon$, and $(x^{k1}, \dots, x^{km}) \in \bar{V}$. Then $t_k \ge \min\{1, \gamma\epsilon/3\bar{\kappa}\}$, where $\bar{\kappa}$ is the Lipschitz constant of f on $B(\bar{x}, 2\epsilon)$.*

Proof. In the proof of Lemma 3.2(ii), assuming $t_k < \min\{1, \gamma\epsilon/3\bar{\kappa}\}$, use $d^k := -g^k$ to get $\langle v^k, g^k\rangle \le \beta|g^k|^2$ as before. Since $\gamma^{-1}t_k|d^k| < \epsilon/3$ yields $v^k \in G_\epsilon(\bar{x})$ as before, note that $|d^k| = |g^k| \le \bar{\kappa}$, since $|x^k - \bar{x}| \le \epsilon/3$ implies $g^k \in G_\epsilon(x^k) \subset G_{1.5\epsilon}(\bar{x})$ and hence $|g^k| \le \bar{\kappa}$. □

With the above replacements, the proofs of section 3 are modified in obvious ways. For instance, in the proof of Theorem 3.3, using (4.3a), we can replace (3.5) by

$$\sum_{k=1}^{\infty} t_k |g^k|^2 < \infty, \tag{4.4}$$

and in its part (ii) we can consider $t_k < \min\{1, \gamma\epsilon/3\bar{\kappa}\}$. In effect, Theorems 3.3–3.6 hold for this variant as well.

Although $d^k := -g^k$ may be better than $d^k := -g^k/|g^k|$ asymptotically, it can be worse initially when $|g^k|$ is still "large" (this, of course, depends on problem scaling). In general, we may wish to scale d^k so that the first trial point $x^k + d^k$ is at a "reasonable" distance from x^k; using the sampling radius ϵ_k as this distance gives the variant analyzed below.

4.2. Searching within the trust region. To restrict the Armijo line search to the sampled trust region $B(x^k, \epsilon_k)$, suppose relations (2.4)–(2.6) in Algorithm 2.1 are replaced by

$$d^k := -\epsilon_k g^k/|g^k|, \tag{4.5}$$

$$t_k := \max\{\, t : f(x^k + td^k) < f(x^k) - \beta t\epsilon_k|g^k|, t \in \{1, \gamma, \gamma^2, \dots\}\,\}, \tag{4.6}$$

$$f(x^{k+1}) < f(x^k) - \beta t_k \epsilon_k |g^k|, \tag{4.7a}$$

$$|x^k + t_k d^k - x^{k+1}| \le \min\{t_k, \epsilon_k\}|d^k|. \tag{4.7b}$$

Then (2.7) still holds, since $|x^{k+1} - x^k| \le 2t_k|d^k| = 2t_k\epsilon_k$. Lemma 3.2(ii) is replaced by the following.

LEMMA 4.2. *Let $\epsilon > 0$ and $\bar{x} \in \mathbb{R}^n$. Assuming $0 \notin G_\epsilon(\bar{x})$, pick $\delta > 0$ as in Lemma* 3.1 *for $C := G_\epsilon(\bar{x})$, and then τ and $\bar{V}$ as in Lemma* 3.2(i). *Suppose $x^k \in B(\bar{x}, \min\{\tau, \epsilon/3\})$, $\epsilon_k = \epsilon$, and $(x^{k1}, \dots, x^{km}) \in \bar{V}$. Then $t_k \geq \gamma/3$.*

Proof. In the proof of Lemma 3.2(ii), for $t_k < \gamma/3$, use $d^k := -\epsilon_k g^k/|g^k|$ to get $\langle v^k, g^k \rangle \leq \beta |g^k|^2$ as before, and then $v^k \in G_\epsilon(\bar{x})$ from $\gamma^{-1} t_k |d^k| < \epsilon/3$ with $|d^k| = \epsilon_k = \epsilon$. ∎

As in section 4.1, we deduce that Theorems 3.3–3.6 hold for this variant as well, since in the proof of Theorem 3.3, using (4.7a), we can replace (3.5) by

$$\sum_{k=1}^{\infty} t_k \epsilon_k |g^k| < \infty. \tag{4.8}$$

4.3. Limiting the line search. Note that relations (2.4)–(2.6), (4.1)–(4.3), and (4.5)–(4.7) have the form

$$d^k := -\alpha_k g^k \quad \text{with} \quad \alpha_k > 0, \tag{4.9}$$

$$t_k := \max\{\, t : f(x^k + t d^k) < f(x^k) - \beta t |d^k||g^k|, t \in \{1, \gamma, \gamma^2, \dots\}\,\}, \tag{4.10}$$

$$f(x^{k+1}) < f(x^k) - \beta t_k |d^k||g^k|, \tag{4.11a}$$

$$|x^k + t_k d^k - x^{k+1}| \leq \min\{t_k, \epsilon_k\} |d^k|, \tag{4.11b}$$

where $\alpha_k := 1/|g^k|$ in sections 2–3, $\alpha_k := 1$ in section 4.1, and $\alpha_k := \epsilon_k/|g^k|$ in section 4.2. The corresponding lower bounds on t_k produced by Lemmas 3.2(ii), 4.1, and 4.2 have the form $t_k \geq \min\{1, \gamma\epsilon/3|d^k|\}$. Procedure 4.3 tests only step sizes that satisfy this bound. It finds $t_k \geq \min\{1, \gamma\epsilon/3|d^k|\}$ when the search direction is good enough (see Lemma 4.4). Otherwise, a *null step* with $t_k := 0$ occurs; then Step 1 resamples (most of) the gradient bundle G_k, so that eventually the search direction is improved sufficiently (unless x^k is already stationary). It will be seen that this null step/resampling mechanism obviates the need for the iterates to be in the set D where f is differentiable.

PROCEDURE 4.3 (limited Armijo line search).

(i) Choose an initial step size $t \geq \min\{1, \gamma\epsilon_k/3|d^k|\}$.
(ii) If $f(x^k + t d^k) < f(x^k) - \beta t |d^k||g^k|$, return $t_k := t$.
(iii) If $t \leq \min\{1/\gamma, \epsilon_k/3|d^k|\}$, return $t_k := 0$.
(iv) Set $t := \gamma t$ and go to (ii).

LEMMA 4.4. *Let $\epsilon > 0$ and $\bar{x} \in \mathbb{R}^n$. Assuming $0 \notin G_\epsilon(\bar{x})$, pick $\delta > 0$ as in Lemma* 3.1 *for $C := G_\epsilon(\bar{x})$, and then τ and $\bar{V}$ as in Lemma* 3.2(i). *Suppose $x^k \in B(\bar{x}, \min\{\tau, \epsilon/3\})$, $\epsilon_k = \epsilon$, $(x^{k1}, \dots, x^{km}) \in \bar{V}$, and $d^k := -\alpha_k g^k$ with $\alpha_k > 0$. Then Procedure* 4.3 *finds a step size $t_k \geq \min\{1, \gamma\epsilon/3|d^k|\}$, and the conclusions of Lemmas* 3.2(ii), 4.1, *and* 4.2 *hold for $\alpha_k = 1/|g^k|$, 1, and $\epsilon_k/|g^k|$, respectively.*

Proof. As in the proof of Lemma 3.2(ii), using relation (3.4) and the form of $d^k := -\alpha_k g^k$, we obtain $\langle v, d^k \rangle < -\beta |d^k||g^k|$ for all $v \in G_\epsilon(\bar{x})$. Let $t \in (0, \epsilon/3|d^k|]$. By Lebourg's mean value theorem, $f(x^k + t d^k) - f(x^k) = t\langle v, d^k \rangle$ for some $v \in \bar{\partial} f(x)$ with $x \in [x^k + t d^k, x^k]$. Then $t|d^k| \leq \epsilon/3$ and $|x^k - \bar{x}| \leq \epsilon/3$ imply $x \in B(\bar{x}, 2\epsilon/3)$ and hence $v \in G_\epsilon(\bar{x})$. Therefore, $f(x^k + t d^k) < f(x^k) - \beta t |d^k||g^k|$ for all $t \in (0, \epsilon/3|d^k|]$, and the conclusion follows from the rules of Procedure 4.3. ∎

Remark 4.5.

(i) We conclude from Lemma 4.4 that Theorems 3.3–3.6 remain valid for step sizes t_k produced by Procedure 4.3 instead of the standard Armijo searches (2.5), (4.2), and (4.6). This follows easily from the proofs of section 3 and the remarks in sections 4.1–4.2.

(ii) The number of f-evaluations made by Procedure 4.3 can be controlled via the choice of the initial step size t at step (i). For instance, if $t := \min\{1, \epsilon_k/3|d^k|\}$, then only one evaluation occurs, and the procedure returns either $t_k := t$ or $t_k := 0$. If the initial step size t looks "too small," e.g., $f(x^k + td^k) < f(x^k) - 0.5t|d^k||g^k|$, we can try expansion by setting $t := t/\gamma$ until $f(x^k + td^k) \geq f(x^k) - \beta t|d^k||g^k|$, in which case $t_k := \gamma t$ is returned. Further, step (iii) of Procedure 4.3 can use a smaller threshold $0 < \underline{t} < \min\{1/\gamma, \epsilon_k/3|d^k|\}$, returning $t_k := 0$ if $t \leq \underline{t}$. Alternatively, the stopping criterion of step (iii) can be ignored until a given number of f-evaluations is reached. Such variations do not impair Lemma 4.4. Note that the theoretical guidelines above leave much freedom for implementations. For instance, choosing a smaller $\underline{t}$ involves the trade-off between the cost of additional f-evaluations during the line search versus the cost of evaluating m gradients at Step 1. On the other hand, using a positive $\underline{t}$ is essential in practice because otherwise an infinite loop might occur due to rounding errors.

(iii) Once Procedure 4.3 replaces the standard Armijo searches (2.5), (4.2), and (4.6), there is no longer any need for keeping x^k in D and including $\nabla f(x^k)$ in G_k at Step 1. This leads to the following simplified variant of Algorithm 2.1. At Step 0, select any $x^1 \in \mathbb{R}^n$. At Step 1, set $G_k := \mathrm{co}\{\nabla f(x^{ki})\}_{i=1}^m$. At Step 5, find t_k via Procedure 4.3. Finally, at Step 6, set $x^{k+1} := x^k + t_k d^k$. Then the requirements of (4.11) are met if $t_k > 0$, whereas the key inequality (2.7) always holds. In effect, Theorems 3.3–3.6 remain valid for this variant. Further, Theorems 3.3–3.6 still hold if the differentiability check of Step 1 is skipped, since $\{x^{ki}\}_{i=1}^m \subset D$ with probability 1. In other words, we may skip the differentiability check of Steps 1 and 6 as in the implementation of [BLO05, section 4], assuming that the user provides reasonable replacements for the gradient at points where it is not defined or that such points are not encountered. In this setting, we may also include $\nabla f(x^k)$ in G_k: although we can not show that the event $x^k \notin D$ has probability zero, it is unlikely to occur in practice.

Acknowledgments. I would like to thank the Associate Editor and the two anonymous referees for their help in improving the exposition of the main results of this paper.

REFERENCES

[BHLO06] J. V. Burke, D. Henrion, A. S. Lewis, and M. L. Overton, *Stabilization via nonsmooth, nonconvex optimization*, IEEE Trans. Automat. Control, 51 (2006), pp. 1760–1769.

[BLO02a] J. V. Burke, A. S. Lewis, and M. L. Overton, *Approximating subdifferentials by random sampling of gradients*, Math. Oper. Res., 27 (2002), pp. 567–584.

[BLO02b] J. V. Burke, A. S. Lewis, and M. L. Overton, *Two numerical methods for optimizing matrix stability*, Linear Algebra Appl., 351/352 (2002), pp. 147–184.

[BLO04] J. V. Burke, A. S. Lewis, and M. L. Overton, *Pseudospectral components and the distance to uncontrollability*, SIAM J. Matrix Anal. Appl., 26 (2004), pp. 350–361.

[BLO05] J. V. Burke, A. S. Lewis, and M. L. Overton, *A robust gradient sampling algorithm for nonsmooth, nonconvex optimization*, SIAM J. Optim., 15 (2005), pp. 751–779.

[Cla83] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983.

[Gol77] A. A. Goldstein, *Optimization of Lipschitz continuous functions*, Math. Programming, 13 (1977), pp. 14–22.

[Kiw96] K. C. Kiwiel, *Restricted step and Levenberg–Marquardt techniques in proximal bundle methods for nonconvex nondifferentiable optimization*, SIAM J. Optim., 6 (1996), pp. 227–249.

[Lew05] A. S. Lewis, *Local structure and algorithms in nonsmooth optimization*, in Optimization and Applications, F. Jarre, C. Lemaréchal, and J. Zowe, eds., Mathematisches Forschungsinstitut Oberwolfach, Oberwolfach, Germany, 2005, pp. 104–106.

[Roc70] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

SIAM J. OPTIM.
Vol. 18, No. 2, pp. 389–412

CODERIVATIVE ANALYSIS OF QUASI-VARIATIONAL INEQUALITIES WITH APPLICATIONS TO STABILITY AND OPTIMIZATION*

BORIS S. MORDUKHOVICH† AND JIŘÍ V. OUTRATA‡

Abstract. We study equilibrium models governed by parameter-dependent quasi-variational inequalities important from the viewpoint of optimization/equilibrium theory as well as numerous applications. The main focus is on quasi-variational inequalities with parameters entering both single-valued and multivalued parts of the corresponding generalized equations in the sense of Robinson. The main tools of our variational analysis involve coderivatives of solution maps to quasi-variational inequalities, which allow us to obtain efficient conditions for robust Lipschitzian stability of quasi-variational inequalities and also to derive new necessary optimality conditions for mathematical programs with quasi-variational constraints. To conduct this analysis, we develop new results on coderivative calculus for structural settings involved in our models. The results obtained are illustrated by applications to some optimization and equilibrium models related to parameterized Nash games of two players and to oligopolistic market equilibria.

Key words. variational analysis, quasi-variational inequalities, equilibrium constraints, parametric optimization, Lipschitzian stability, generalized differentiation

AMS subject classifications. 49J52, 49K40, 58C20

DOI. 10.1137/060665609

1. Introduction.

This paper is devoted to the study of optimization and equilibrium problems involving the so-called parameterized *quasi-variational inequalities* (QVIs) of the following type: given a *parameter* $x \in \mathbb{R}^n$, find a *decision* vector $y \in \Gamma(x,y) \subset \mathbb{R}^m$ such that

$$\langle g(x,y), v-y\rangle \geq 0 \ \text{ for all } \ v \in \Gamma(x,y), \tag{1.1}$$

where $g\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ is a *single-valued* continuously differentiable vector function, while $\Gamma\colon \mathbb{R}^n \times \mathbb{R}^m \rightrightarrows \mathbb{R}^m$ is a *set-valued* mapping (multifunction) between the corresponding finite-dimensional spaces. We always assume in this paper that the mapping Γ is of *closed graph* and takes *convex values* $\Gamma(x,y)$.

QVIs were introduced by Bensoussan and Lions in a series of papers (see, e.g., [3]) in connection with *impulse optimal control* problems. They have been extensively studied in numerous publications, mainly from the viewpoints of existence of solutions and numerical methods; cf. [2, 5, 7, 10, 19, 29], among others. Besides the original motivation, models in the form of QVIs and their special subclass, *implicit complementarity problems*, were particularly used, e.g., in

- *continuum mechanics* (filtration through porous media [2], contact problems with compliant obstacles [33], contact problems with Coulomb friction [15, 4]);

*Received by the editors July 20, 2006; accepted for publication (in revised form) November 10, 2006; published electronically May 7, 2007.
http://www.siam.org/journals/siopt/18-2/66560.html

†Department of Mathematics, Wayne State University, Detroit, MI 48202 (boris@math.wayne.edu). The research of this author was partly supported by the National Science Foundation under grants DMS-0304989 and DMS-0603846 and by the Australian Research Council under grant DP-0451168.

‡Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, 18208 Prague, Czech Republic (outrata@utia.cas.cz). The research of this author was supported by grant A 107 5402 of the Grant Agency of the Academy of Sciences of the Czech Republic.

- *economics* (noncooperative games [14, 11], oligopolistic markets [33], and network and traffic equilibria [7]); and
- *biology* (competition between different species or within a species [13]).

Much less attention has been paid to the study of *parameter-dependent* QVIs, especially those where *both* mappings g and Γ in (1.1) depend on parameters. The primary goal of this paper is to undertake such a study concentrating mainly on *sensitivity/stability* of solution maps to (1.1) defined by

$$S(x) := \big\{y \in \mathbb{R}^m \big| \ \langle g(x,y), v-y\rangle \geq 0 \ \text{ whenever } \ v \in \Gamma(x,y)\big\}, \quad x \in \mathbb{R}^n,$$

and *necessary optimality conditions* for local optimal solutions to mathematical programs with *equilibrium constraints* governed by QVIs of type (1.1).

Let us mention previous developments and results known in these directions for QVIs. The papers [18, 31] concern stability issues for QVIs and contain conditions ensuring the existence of a single-valued Lipschitzian localization of $S(\cdot)$ around a reference point. This localization is then described via generalized Jacobians [6] for Lipschitzian mappings. The same type of analysis can also be found in [4], where the authors consider a parameterized QVI describing a contact problem with Coulomb friction. QVIs typically have, however, rather complicated solution sets, and so the results of [4, 18, 31] can be applied only in problems of a special kind. A general study of sensitivity and optimality aspects for parameterized variational systems is given in [27], but the results obtained therein are not specified for QVIs and do not imply those derived in this paper. It is worth emphasizing that QVIs, even in rather simple settings, are significantly different from standard variational inequalities and complementarity problems, demanding therefore new devices and tools for their variational analysis. One of the most crucial characteristic features of QVIs is their *intrinsic nonsmoothness*, which unavoidably requires the usage of appropriate tools of generalized differentiation. In this paper we apply to the study of QVIs the generalized differential constructions and their calculus mainly developed by the first author (see, e.g., [22, 24, 26] and also [37] and the references therein). However, for our purposes we need new calculus results of generalized differentiation specified for applications to QVIs, which are developed in what follows.

Using the standard definition of the normal cone to convex sets, we can rewrite the QVI (1.1) in Robinson's form of the *generalized equation* (GE)

$$0 \in g(x,y) + N_{\Gamma(x,y)}(y), \quad y \in \Gamma(x,y), \tag{1.2}$$

where $N_{\Gamma(x,y)}(y)$ stands for the usual normal cone to the set $\Gamma(x,y)$ at y. Note that (1.2) is different from the conventional form of GEs introduced in [34] for the case of constant convex sets $\Gamma(x,y) \equiv \Gamma$ when (1.2) reduces to the classical variational inequality. The case of variable sets that is characteristic for QVIs (even with no dependence on parameters) happens to be significantly more involved.

The *solution map* $S(x)$ to the QVI (1.1) written in the GE form (1.2) is given by

$$S(x) = \big\{y \in \mathbb{R}^m \big| \ 0 \in g(x,y) + N_{\Gamma(x,y)}(y)\big\}. \tag{1.3}$$

One of the primary goals of local sensitivity analysis conducted in this paper is to find verifiable conditions ensuring *robust Lipschitzian stability* of the solution map (1.3) with respect to parameter perturbations. This can be done on the basic of the *coderivative criterion* for robust Lipschitzian behavior of multifunctions established in [22, 23] via the *coderivative* construction for set-valued mappings that reduces to the *adjoint* derivative operator in the smooth single-valued case.

In order to apply this criterion to QVIs, we need to derive efficient *upper estimates* of the coderivative for the solution map (1.3). This is one of the major technical achievements of the paper, obtained on the base of new rules of *coderivative calculus*. Furthermore, the established coderivative estimates are employed in the paper for deriving *new optimality conditions* for mathematical programs with *QVI constraints*. The results obtained are illustrated by applications to certain optimization and equilibrium models related to *parameterized Nash games* of two players and to *oligopolistic market equilibria*.

The *outline* of the paper is as follows. In section 2 we present and briefly discuss some basic definitions and preliminaries from *variational analysis* and *generalized differentiation* needed for formulating the main results. Section 3 is devoted to the study of *coderivatives* of the multivalued term in (1.2), which is done by deriving new coderivative *calculus rules* that take into account the specific *amenable* structure of the composition in (1.2). Observe that the obtained calculus rules involve new *calmness* assumptions on multifunctions, which are not conventional for subdifferential and coderivative calculus and are significantly weaker than those in known results. In section 4 we apply the new calculus rules to establish efficient *upper estimates* of the *coderivative* for the *solution map* (1.3) to the QVI (1.1) derived under appropriate calmness assumptions and *constraint qualifications*.

Section 5 is devoted to deriving verifiable conditions ensuring *robust Lipschitzian stability* of solutions maps to QVIs established on the base on the aforementioned coderivative criterion and calculus rules. We also present new *upper estimates* of the *exact Lipschitzian bound* for solution maps, which is certainly of interest for both qualitative and numerical analysis. The results obtained are illustrated by establishing Lipschitzian stability of the QVI corresponding to the parameterized *Nash games* studied in [14].

In the concluding section 6 we consider a class of *mathematical programs with QVI constraints*, which can be treated as a subclass of *mathematical programs with equilibrium constraints* (MPECs) intensively studied and used in modern optimization theory and its numerous applications; see, e.g., the books [10, 21, 27, 33] and the references therein. However, we are not familiar with any results that can be applied to optimization problems with equilibrium constraints governed by the QVIs (1.1) under consideration. The new *necessary optimality conditions* for such problems obtained in section 6 strongly employ the coderivative analysis developed in sections 3 and 4. The results obtained are illustrated by the applications to an MPEC with QVI constraints corresponding to the parameterized *Nash games* considered in section 5 and to an *oligopolistic market equilibrium model* in the vein of [11, 30].

Our *notation* is basically standard; see, e.g., [26, 37]. Recall that x^i stands for the ith component of a vector $x \in \mathbb{R}^n$ always considered as a vector-column, E is the unit matrix, $\mathbb{B}$ is the closed unit ball of the space in question, $\mathbb{N} := \{1, 2, \dots\}$, and $A^T (= A^*)$ signifies the transposition of (the adjoint operator to) the matrix A with the vector-row x^T corresponding to the vector-column $x \in \mathbb{R}^n$; thus $x^T y = \langle x, y \rangle$ connotes the inner product of the vectors x and y. Given a differentiable vector function $f\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^l$ of two variables, say x and y, the symbol $\nabla_x f(x, y)$ stands for the partial Jacobian (or gradient in the scalar case of $l = 1$) with respect to x. In a special case of $x = y$, we use the notation $\nabla_1 f(y, y)$ to prevent confusion. Similarly, if $\varphi\colon \mathbb{R}^n \times \mathbb{R}^m \to \overline{\mathbb{R}} := (-\infty, \infty]$ is a nondifferentiable extended-real-valued function of x and y, then $\partial_x \varphi(x, y)$ signifies its partial subdifferential with respect to x. As usual, $\delta(x; \Omega) = \delta_\Omega(x)$ denotes the *indicator function* of the set Ω equal 0 for $x \in \Omega$

and ∞ for $x \notin \Omega$. The graph of a set-valued mapping $F\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is

$$\operatorname{gph} F := \big\{(x,y) \in \mathbb{R}^n \times \mathbb{R}^m \big|\ y \in F(x)\big\}.$$

2. Some concepts and tools of variational analysis. In this section we review certain preliminary material from *variational analysis* and *generalized differentiation* that is extensively used throughout the paper; see [22, 24, 26, 37] for more details and references. Given a set $\Omega \subset \mathbb{R}^n$ locally closed around $\bar{x} \in \Omega$, define the (basic, limiting) *normal cone* to Ω at $\bar{x}$ by

$$N_\Omega(\bar{x}) = N(\bar{x};\Omega) := \Big\{v \in \Omega \ \Big|\ \exists\, x_k \to \bar{x},\ \exists\, w_k \in \Pi_\Omega(x_k),\ \exists\, \alpha_k \ge 0 \ \text{ with } \ \alpha_k(x_k - w_k) \to v \ \text{ as } \ k \to \infty\Big\}, \tag{2.1}$$

where $\Pi_\Omega(x)$ stands for the *Euclidean projector* of x on Ω. There are several equivalent representations of the normal cone that can be found in the aforementioned references. For convex sets Ω, the normal cone (2.1) reduces to the normal cone of convex analysis, while it is generally *nonconvex* even for simple sets on the plane, e.g., for $\Omega = \operatorname{gph}|x|$ at $0 \in \mathbb{R}^2$. Moreover, the *convexification* (taking the closed convex hull) of N_Ω, which reduces to Clarke's normal cone [6], often gives the *whole space* (as for $\Omega = \operatorname{gph}|x|$) or at least a *linear subspace* of the maximal dimension. In particular, this always happens for the so-called *graphically Lipschitzian sets* around $\bar{x}$, i.e., those which are locally homeomorphic to the graph of a locally Lipschitzian vector function. The latter is automatically the case for graphs of *monotone operators* and *subdifferential mappings* generated by "nice" (e.g., convex, saddle, amenable, etc.) functions; see [26, 37] for precise results and discussions.

The normal cone (2.1) to graphical sets generates the following derivative-like construction for set-valued (and single-valued) mappings that plays the crucial role in the variational analysis for QVIs conducted in this paper. Given a set-valued mapping $F\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^m$, we define its *coderivative* at $(\bar{x},\bar{y}) \in \operatorname{gph} F$ as a positive homogeneous multifunction $D^*F(\bar{x},\bar{y})\colon \mathbb{R}^m \rightrightarrows \mathbb{R}^n$ with the values

$$D^*F(\bar{x},\bar{y})(u) = \big\{v \in \mathbb{R}^n \big|\ (v,-u) \in N\big((\bar{x},\bar{y});\operatorname{gph} F\big)\big\}, \tag{2.2}$$

where $\bar{y} = f(\bar{x})$ is omitted if $F = f\colon \mathbb{R}^n \to \mathbb{R}^m$ is single-valued. If $f\colon \mathbb{R}^n \to \mathbb{R}^m$ is *smooth* (continuously differentiable) around $\bar{x}$, then

$$D^*f(\bar{x})(u) = \big\{\nabla f(\bar{x})^T u\big\} \ \text{ for all } \ u \in \mathbb{R}^m; \tag{2.3}$$

the latter holds when f is merely *strictly differentiable* at $\bar{x}$.

Given an extended-real-valued function $\varphi\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ finite at $\bar{x}$ and lower semicontinuous around this point, define the (basic, limiting) *subdifferential* of φ at $\bar{x}$ by

$$\partial\varphi(\bar{x}) := \big\{v \in \mathbb{R}^n \big|\ (v,-1) \in N((\bar{x},\varphi(\bar{x}));\operatorname{epi}\varphi)\big\} \tag{2.4}$$

via the normal cone (2.1) to the epigraph

$$\operatorname{epi}\varphi := \big\{(x,\mu) \in \mathbb{R}^{n+1} \big|\ \mu \ge \varphi(x)\big\}.$$

There are well-known analytic representations of $\partial\varphi(\bar{x})$ via limits of Fréchet, proximal, viscosity subgradients, etc. Note the useful *scalarization formula*

$$D^*f(\bar{x})(u) = \partial\langle u, f\rangle(\bar{x}) \ \text{ whenever } \ u \in \mathbb{R}^m,$$

which relates the coderivative (2.2) of the mapping $f\colon \mathbb{R}^n \to \mathbb{R}^m$ locally Lipschitzian around $\bar{x}$ with the subdifferential (2.4) of its Lagrange scalarization

$$\langle u, f\rangle(x) := \langle u, f(x)\rangle, \quad u \in \mathbb{R}^m, \ \ x \in \mathbb{R}^n.$$

Observe also the subdifferential representation

$$N(\bar{x};\Omega) = \partial\delta(\bar{x};\Omega), \quad \bar{x} \in \Omega, \tag{2.5}$$

of the normal cone (2.1) via the subdifferential of the set indicator function.

In spite of (actually due to) the *nonconvexity* of our basic generalized differential constructions (2.1), (2.2), and (2.4), they possess *full calculus* (various rules for intersections of sets, compositions of mappings, mean and marginal values of functions, etc.), which happens to be significantly better than for their convex-valued counterparts. The fundamental fact behind this generalized differential calculus is the *extremal principle* of variational analysis that can be treated as a *variational* counterpart of the classical separation theorem in nonconvex settings and plays a crucial role in variational analysis in the absence (and also in the presence) of convexity; see [26, 27] for all the details.

As mentioned, one of the primary goals of the paper is to provide a local *sensitivity analysis* for solution maps to QVIs. In what follows, we focus our attention on *robust Lipschitzian stability* generated by the so-called Aubin's Lipschitz-like (or "pseudo-Lipschitzian" [1]) property, which happens to be the most natural extension of the classical local Lipschitz continuity to set-valued mappings; see [26, 37] for more discussions. Recall that $F\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ has the (Aubin) *Lipschitz-like property* around $(\bar{x},\bar{y}) \in \operatorname{gph} F$ with modulus $\ell \geq 0$ if there are neighborhoods U of $\bar{x}$ and V of $\bar{y}$ such that

$$F(x) \cap V \subset F(u) + \ell\|x-u\|\mathbb{B} \ \text{ whenever } \ x,u \in U. \tag{2.6}$$

The infimum of all the moduli ℓ, for which (2.6) holds with some neighborhoods U and V, is called the *exact Lipschitzian bound* of F around $(\bar{x},\bar{y})$ and is denoted by $\operatorname{lip} F(\bar{x},\bar{y})$.

When $V = \mathbb{R}^m$ in (2.6), this property reduces to the (Hausdorff) *local Lipschitzian property* of F around the point of the *domain*

$$\bar{x} \in \operatorname{dom} F := \{x \in \mathbb{R}^n |\ F(x) \neq \emptyset\}$$

extending the classical local Lipschitzian behavior to the case of set-valued mappings. From this viewpoint, the main difference between Aubin's and Hausdorff's Lipschitzian properties is that the former is a localization of the latter around the point of the *graph* $(\bar{x},\bar{y}) \in \operatorname{gph} F$ versus that of merely the domain $\bar{x} \in \operatorname{dom} F$. The graph localization allows us to study efficiently Lipschitzian stability of *unbounded* multifunctions, which cannot be covered by the Hausdorff property. Furthermore, the Aubin property (2.6) of F around $(\bar{x},\bar{y})$ happens to be *equivalent* to the well-recognized *metric regularity* and *linear openness* properties of the inverse mapping F^{-1} around $(\bar{y},\bar{x})$; see [26, 37].

A great advantage of the coderivative (2.2) is the possibility to *fully characterize* in its terms the robust Lipschitzian (and metric regularity/linear openness) behavior of set-valued mappings that has been done in [22, 23]. The following *coderivative criterion* (or Mordukhovich criterion as in [37, Theorem 9.40]) holds true: a mapping

$F\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ locally closed-graph around $(\bar{x},\bar{y})$ is *Lipschitz-like* around this point *if and only if*

$$D^*F(\bar{x},\bar{y})(0)=\{0\}. \tag{2.7}$$

Moreover, the *exact Lipschitzian bound* of F around $(\bar{x},\bar{y})$ is computed by

$$\operatorname{lip} F(\bar{x},\bar{y})=\|D^*F(\bar{x},\bar{y})\|=\sup\Big\{\|v\|\;\Big|\;v\in D^*F(\bar{x},\bar{y})(u),\ \|u\|\le 1\Big\} \tag{2.8}$$

via the *norm* of $D^*F(\bar{x},\bar{y})\colon \mathbb{R}^m \rightrightarrows \mathbb{R}^n$ as a positive homogeneous mapping.

In this paper we extensively use one more property of set-valued mappings, which is weaker than (2.6) and relates to linear rate/Lipschitzian behavior of F *at* the point in question. Following [37], we say that $F\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is *calm* at $(\bar{x},\bar{y})\in \operatorname{gph} F$ with modulus $\ell\ge 0$ if there are neighborhoods U of $\bar{x}$ and V of $\bar{y}$ such that

$$F(x)\cap V\subset F(\bar{x})+\ell\|x-\bar{x}\|\mathbb{B}\ \ \text{for all}\ \ x\in U. \tag{2.9}$$

If $V=\mathbb{R}^m$ in (2.9), then the calmness property defined in (2.9) reduces to Robinson's *upper Lipschitzian property* of F *at* $\bar{x}\in\operatorname{dom} F$; see [34].

The principal difference between properties (2.6) and (2.9) is that the former involves all *pairs* of independent domain vectors (x,u) around $\bar{x}$, while the latter *fixes* $u=\bar{x}$. By this, the calmness property *does not* get back the classical local Lipschitzian behavior for single-valued mappings; furthermore, it does not exclude that $F(x)\ne\emptyset$ for x near $\bar{x}$. On the other hand, the calmness/upper Lipschitzian property always holds [35] for *piecewise polyhedral* set-valued mappings, i.e., those whose graph can be represented as a union of finitely many convex polyhedral sets.

3. New rules of coderivative calculus. Throughout the paper, we assume that the set-valued mapping Γ generating the QVI in (1.1) admits the *representation*

$$\Gamma(x,y):=\{z\in\mathbb{R}^m|\ q(x,y,z)\in\Theta\}, \tag{3.1}$$

where $q\colon \mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}^m\to\mathbb{R}^s$ is a vector function *twice continuously differentiable* around the points in question, and where Θ is a *closed convex* subset of $\mathbb{R}^s$. Additionally, q and Θ have to satisfy certain requirements ensuring that Γ in (3.1) is *convex-valued*, which is essential to ensure the *strong amenable* structure in the representation of $N_\Gamma(x,y)(z)$; see below. The convex-valuedness property of $\Gamma(x,y)$ holds, e.g., if Θ is a convex cone with vertex at 0 and if $q(x,y,\cdot)$ is Θ-*convex* (in the standard sense) for all $(x,y)\in\mathbb{R}^n\times\mathbb{R}^m$. Furthermore, we impose the basic *constraint qualification* (CQ) condition:

$$\left.\begin{aligned}(\nabla q(\bar{x},\bar{y},\bar{y}))^T u=0\\ u\in N_\Theta(q(\bar{x},\bar{y},\bar{y}))\end{aligned}\right\}\Longrightarrow u=0. \tag{3.2}$$

The main goal of this section is to study the *coderivative* (2.2) of the set-valued mapping $(x,y)\rightrightarrows N_{\Gamma(x,y)}(y)$ in the generalized equation form (1.2) of the QVI under consideration. As mentioned in section 1, the coderivative of this mapping plays a significant role in what follows. We intend to derive efficient *upper estimates* of this coderivative for Γ defined in (3.1), which provide new rules of coderivative calculus for set-valued mappings of a special structure that is characteristic for QVIs.

It follows from formula (2.5) and the structure of Γ in (3.1) that $N_{\Gamma(x,y)}$ admits the *composite subdifferential representation*

$$N_{\Gamma(x,y)}(z) = \partial_z \psi(x,y,z) \;\text{ with }\; \psi := \delta_\Theta \circ q.$$

Since the basic CQ (3.2) is persistent in a neighborhood, by the *robustness* (closed-graph) property of the normal cone (2.1), the composite function ψ is *strongly amenable* around $(\bar{x},\bar{y},\bar{y})$ in the sense of [37], and we have by [37, Theorem 10.49] that

$$N_{\Gamma(x,y)}(z) = (\nabla_z q(x,y,z))^T N_\Theta(q(x,y,z))$$

whenever (x,y,z) is sufficiently close to $(\bar{x},\bar{y},\bar{y})$. This means that, for the purpose of our local analysis, we can replace the GE (1.2) by

$$0 \in g(x,y) + (\nabla_3 q(x,y,y))^T N_\Theta(q(x,y,y)) \tag{3.3}$$

considered for all (x,y,z) around $(\bar{x},\bar{y},\bar{y})$, where $\nabla_3 q(x,y,y)$ signifies, according to our notation in section 1, the partial derivative of $q(x,y,z)$ with respect to z at (x,y,y).

In this section we focus our attention on deriving upper estimates for the coderivative of the multivalued term in (3.3) denoted by

$$Q(x,y) := (\nabla_3 q(x,y,y))^T N_\Theta(q(x,y,y)). \tag{3.4}$$

This set-valued mapping $Q\colon \mathbb{R}^n \times \mathbb{R}^m \rightrightarrows \mathbb{R}^m$ is *closed-graph* around $(\bar{x},\bar{y},\bar{y})$ due to the robustness of the normal cone and the continuity of q. To unburden the notation, put

$$r(x,y) := q(x,y,y) \;\text{ and }\; p(x,y) := \nabla_3 q(x,y,y). \tag{3.5}$$

We clearly have the representation

$$\nabla_y r(x,y) = \nabla_2 q(x,y,y) + \nabla_3 q(x,y,y).$$

The next theorem on *coderivative calculus* for multifunction of the special type (3.4) consists of two parts providing the corresponding *upper estimates* for the coderivative $D^*Q(\bar{x},\bar{y},\bar{v})$ of the set-valued term Q in (3.3) with some $\bar{v} \in Q(\bar{x},\bar{y})$. The first estimate valid under a *partial* modification of the basic *first-order* qualification condition (3.2) ends up with an expression containing the coderivative of the composition $N_\Theta \circ r$, while the second one elaborates the coderivative of the latter composition under an additional *calmness* assumption, which plays a role of some *second-order qualification condition*.

THEOREM 3.1 (coderivative calculus for special compositions). *Under the standing assumptions above, suppose that the basic CQ* (3.2) *is strengthened as*

$$\left.\begin{aligned}(p(\bar{x},\bar{y}))^T u = 0\\ u \in N_\Theta(r(\bar{x},\bar{y}))\end{aligned}\right\} \Longrightarrow u = 0, \tag{3.6}$$

and let $\bar{v} \in Q(\bar{x},\bar{y})$. *Then the following assertions hold:*

(i) *For all* $u \in \mathbb{R}^m$ *we have the coderivative upper estimate*

$$D^*Q(\bar{x},\bar{y},\bar{v})(u) \subset \bigcup_{\substack{d \in N_\Theta(r(\bar{x},\bar{y}))\\ (p(\bar{x},\bar{y}))^T d = \bar{v}}} \Big[(\nabla_{x,y}((p(\bar{x},\bar{y}))^T d))^T u + D^*(N_\Theta \circ r)(\bar{x},\bar{y},d)(p(\bar{x},\bar{y})u)\Big].$$

(ii) *Define the set-valued mapping* $M\colon R^s \times \mathbb{R}^s \rightrightarrows \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^s$ *by*

$$M(\vartheta) := \left\{(x,y,d) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^s \,\middle|\, \begin{bmatrix} r(x,y) \\ d \end{bmatrix} + \vartheta \in \operatorname{gph} N_\Theta \right\}, \tag{3.7}$$

and assume that it is calm at the points $(0,\bar{x},\bar{y},d)$ *satisfying*

$$d \in N_\Theta(r(\bar{x},\bar{y})) \;\text{ and }\; (p(\bar{x},\bar{y}))^T d = \bar{v}.$$

Then for all $u \in \mathbb{R}^m$ *we have the inclusion*

$$\begin{aligned} D^*Q(\bar{x},\bar{y},\bar{v})(u) \subset \bigcup_{\substack{d \in N_\Theta(r(\bar{x},\bar{y})) \\ (p(\bar{x},\bar{y}))^T d = \bar{v}}} & \Big[(\nabla_{x,y}((p(\bar{x},\bar{y}))^T d))^T u \\ & + (\nabla r(\bar{x},\bar{y}))^T D^* N_\Theta(r(\bar{x},\bar{y}),d)(p(\bar{x},\bar{y})u)\Big]. \end{aligned} \tag{3.8}$$

Proof. To justify assertion (i), we represent the multifunction Q under consideration (3.4) as the *composition*

$$Q(x,y) = (f \circ F)(x,y) \tag{3.9}$$

of the single-valued mapping $f\colon \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^s \to \mathbb{R}^m$ defined by

$$f(c,e,d) := (p(c,e))^T d$$

and the set-valued mapping $F\colon \mathbb{R}^n \times \mathbb{R}^m \rightrightarrows \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^s$ defined by

$$F(x,y) := \begin{bmatrix} x \\ y \\ N_\Theta(r(x,y)) \end{bmatrix}.$$

Observe that the mapping f in (3.9) is single-valued and *smooth* under the assumptions made, while F is always *set-valued*. The *coderivative chain rule* most appropriate in this setting is given in [24, Corollary 5.3]. To apply it, we denote

$$\begin{aligned} G(x,y,v) := F(x,y) \cap f^{-1}(v) = \Big\{(c,e,d) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^s \,\Big|\, & c = x,\ e = y, \\ & d \in N_\Theta(r(x,y)),\ (p(x,y))^T d = v\Big\} \end{aligned}$$

and show that this mapping is *uniformly bounded* around $(\bar{x},\bar{y},\bar{v})$ under the CQ (3.6). Arguing by contradiction, suppose that there are sequences

$$x_k \to \bar{x},\ y_k \to \bar{y},\ d_k \in N_\Theta(r(x_k,y_k)),\ v_k = (p(x_k,y_k))^T d_k, \;\text{ and }\; v_k \to \bar{v}$$

such that $\|d_k\| \ge k$ for all $k \in \mathbb{N}$. By passing to a subsequence if necessary, we find $d \in \mathbb{R}^s$ satisfying the relationships

$$\frac{d_k}{\|d_k\|} \to d \;\text{ as }\; k \to \infty \;\text{ with }\; \|d\| = 1.$$

By continuity of r and p in (3.5) and robustness of the normal cone $N_\Theta(\cdot)$, we get

$$d \in N_\Theta(r(\bar{x},\bar{y})) \;\text{ and }\; (p(\bar{x},\bar{y}))^T d = 0,$$

which contradicts (3.6). Hence, by the aforementioned coderivative chain rule from [24], we arrive at the inclusion

$$D^*Q(\bar{x},\bar{y},\bar{v})(u) \subset \bigcup_{\substack{d\in N_\Theta(r(\bar{x},\bar{y}))\\ (p(\bar{x},\bar{y}))^T d=\bar{v}}} D^*F(\bar{x},\bar{y},\bar{x},\bar{y},d)\circ D^*f(\bar{x},\bar{y},d)(u). \tag{3.10}$$

Furthermore, since p in (3.5) is smooth around $(\bar{x},\bar{y})$ due to the twice continuously differentiability of q, we get by (2.3) that

$$D^*f(\bar{x},\bar{y},d)(u)=\begin{bmatrix}(\nabla_{x,y}((p(\bar{x},\bar{y}))^Td))^T\\ p(\bar{x},\bar{y})\end{bmatrix}u.$$

To justify the coderivative upper estimate in (i), it remains to observe by the above relationships that for all $u=(x^*,y^*,d^*)$ one has the equality

$$D^*F(\bar{x},\bar{y},\bar{x},\bar{y},d)(u)=\begin{bmatrix}x^*\\ y^*\end{bmatrix}+D^*(N_\Theta\circ r)(\bar{x},\bar{y},d)(d^*)$$

due to the coderivative sum rule from [26, Theorem 1.62].

Next we show that the coderivative upper estimate in (i) implies the one in (ii) under the additional *calmness* assumption made in (ii). The difference between these two estimates is that instead of the coderivative $D^*(N_\Theta\circ r)$ of the composition $N_\Theta\circ r$ in (i) we obtain the estimate in (3.8) via the gradient of r and the coderivative of N_Θ *separately*, which is much more convenient for further applications. To proceed, consider the set-valued mapping $T:=N_\Theta\circ r$ and observe that

$$\operatorname{gph}T=\left\{(x,y,d)\in\mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}^s\,\middle|\,\begin{bmatrix}r(x,y)\\ d\end{bmatrix}\in\operatorname{gph}N_\Theta\right\}.$$

Since $\operatorname{gph}T=M(0)$ for the mapping $M(\cdot)$ in (3.7), we apply [16, Theorem 4.1] under the aforementioned calmness assumption and get

$$N_{\operatorname{gph}T}(\bar{x},\bar{y},d)\subset\begin{bmatrix}(\nabla r(\bar{x},\bar{y}))^T & 0\\ 0 & E\end{bmatrix}N_{\operatorname{gph}N_\Theta}(r(\bar{x},\bar{y}),d),$$

which is equivalent to the inclusion

$$D^*(N_\Theta\circ r)(\bar{x},\bar{y},d)(d^*)\subset(\nabla r(\bar{x},\bar{y}))^TD^*N_\Theta(r(\bar{x},\bar{y}),d)(d^*),\quad d^*\in\mathbb{R}^s. \tag{3.11}$$

Substituting (3.11) into the one in (i), we arrive at the coderivative upper estimate (3.8) and complete the proof of the theorem. □

Observe that the expression $D^*N_\Theta=D^*\partial\delta_\Theta$ stands for the *second-order subdifferential* of the indicator function of Θ, according to the general definition of the second-order subdifferential for extended-real-valued functions φ as the coderivative of the first-order subdifferential of φ; see [24, 26, 28] for more details and second-order subdifferential calculus. Hence, the final upper estimate (3.8) of Theorem 3.1 contains *second-order information* on the data involved. Furthermore, the *calmness* assumption in Theorem 3.1(ii) automatically holds, by the *coderivative criterion* (2.7), under the *second-order CQ* condition imposed in the following corollary. Note that the second part of this corollary realizes another possibility to ensure the aforementioned calmness property: the *polyhedral structure* of the initial data of the QVI under consideration.

COROLLARY 3.2 (coderivative calculus under second-order CQ and polyhedrality assumptions). *In addition to first-order qualification condition* (3.6)*, suppose that either*

(a) *the second-order CQ*

$$D^*N_\Theta(r(\bar{x},\bar{y},d)(0)\cap \ker(\nabla r(\bar{x},\bar{y}))^T=\{0\} \tag{3.12}$$

is fulfilled for all $d\in N_\Theta(r(\bar{x},\bar{y}))$ *with* $(p(\bar{x},\bar{y}))^Td=\bar{v}$*; or*

(b) *the set* Θ *in* (3.1) *is polyhedral and the mapping* p *in* (3.5) *is affine.*

Then the coderivative upper estimate (3.8) *is satisfied.*

Proof. To justify (3.8) under the assumptions made, we need to check (by assertion (ii) of Theorem 3.1) that the fulfillment of either (a) or (b) in the corollary implies the calmness requirement of the theorem. As mentioned in section 2, the calmness property of a set-valued mapping at a reference point is automatic when the mapping is *Lipschitz-like* (2.6) around this point. The latter property is *characterized* via the coderivative criterion (2.7). Thus it is sufficient to compute the coderivative of the mapping $M(\cdot)$ given in (3.7). Mappings of this type have been done in [26, Theorem 4.31]. This gives us, by the specific structure of $M(\cdot)$ in (3.7), that the relationship (2.7) is *equivalent* to the second-order CQ (3.12). This justifies the coderivative upper estimate (3.8) in case (a).

To proceed in case (b), recall from section 2 that $M(\cdot)$ in (3.7) is also calm at $(0,\bar{x},\bar{y},d)$ if it is *upper Lipschitzian* at $\vartheta=0$. By [35, Proposition 3], this is always the case for *polyhedral* multifunctions. But the latter property is ensured, due to [35, Proposition 1], by the assumptions made in (b). This completes the proof of the corollary. □

Efficient applications of the coderivative estimate (3.8) largely depend on computing/estimating the second-order term D^*N_Θ. This has been done in various important settings in [9, 17, 20, 27, 28, 32, 38]; see also the references therein.

In numerous situations we have $\Theta=\mathbb{R}^s_-$, the nonpositive orthant of $\mathbb{R}^s$. This case corresponds to *complementarity conditions* and was first analyzed in [9]. To express the coderivative $D^*N_{\mathbb{R}^s_-}(c,d)$ at any point $(c,d)\in \operatorname{gph} N_{\mathbb{R}^s_-}$, consider the *index sets*:

$$\begin{aligned} L(c)&:=\{i\in\{1,\dots,s\}\mid c^i<0\},\\ I_+(d)&:=\{i\in\{1,\dots,s\}\mid d^i>0\},\\ I_0(c,d)&:=\{i\in\{1,\dots,s\}\mid c^i=0,\ d^i=0\}. \end{aligned}$$

Clearly, these sets form a partition of $\{1,\dots,s\}$. Then the *coderivative formula* for precisely computing $D^*N_{\mathbb{R}^s_-}(c,d)(u)$ whenever $u\in\mathbb{R}^s$ is

$$D^*N_{\mathbb{R}^s_-}(c,d)(u)=\left\{v\in\mathbb{R}^s\;\middle|\; v^i\begin{cases}=0 & \text{if}\quad i\in L(c),\text{ or }\ i\in I_0(c,d)\text{ and }u^i<0\\ \in\mathbb{R} & \text{if}\quad i\in I_+(d)\cup\ I_0(c,d)\text{ and }u^i=0\\ \in\mathbb{R}_+ & \text{if}\quad i\in I_0(c,d)\text{ and }\ u^i\ge 0\end{cases}\right\}.$$

Let us employ this formula in the following illustrative example, which is actually a part of some applied models considered below in this paper.

Example 3.3 (computing the coderivative of the normal cone mapping). Consider the mapping $Q(x,y)$ from (3.4) with $n=1,\ m=2,\ s=2,\ \Theta=\mathbb{R}^2_-$, and

$$q^1(x,y,z):=y^1+z^2-15-x,\quad q^2(x,y,z):=y^2+z^1-15-x.$$

Using our notation in (3.5), we have in this case

$$r(x,y)=\begin{bmatrix} y^1+y^2-15-x \\ y^2+y^1-15-x \end{bmatrix} \quad \text{and} \quad p(x,y)=\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Pick the point $(\bar{x},\bar{y})=(0,9,6)$ and put $\bar{v}=(0,1)$. It is easy to see that $d=(1,0)^T$ is the only vector satisfying the equation $(p(\bar{x},\bar{y}))^T d=\bar{v}$. Clearly

$$r(\bar{x},\bar{y})=\begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{and thus} \quad (r(\bar{x},\bar{y}),d)\in \operatorname{gph} N_{\mathbb{R}^2_-}.$$

Furthermore, all the assumptions in case (b) of Corollary 3.2 are satisfied and we have

$$I_+(d)=\{1\}, \quad I_0(r(\bar{x},\bar{y}),d)=\{2\}$$

in the coderivative formula above. Taking into account that the first term on the right-hand side of (3.8) vanishes, we arrive at the coderivative upper estimate

$$D^*Q(\bar{x},\bar{y},\bar{v})(u)$$
$$\subset \left\{ \begin{bmatrix} -v^1-v^2 \\ v^1+v^2 \\ v^1+v^2 \end{bmatrix} \;\middle|\; v^1\in\mathbb{R},\ v^2 \begin{cases} =0 & \text{if} \quad u^1<0 \\ \in\mathbb{R} & \text{if} \quad u^1=0 \\ \in\mathbb{R}_+ & \text{if} \quad u^1>0 \end{cases} \right\} = \{(-a,a,a)\in\mathbb{R}^3 |\ a\in\mathbb{R}\}$$

whenever $u^2=0$. Otherwise $D^*Q(\bar{x},\bar{y},\bar{v})(u)=\emptyset$.

4. Coderivatives of solutions maps to QVIs. The main goal of this section is to derive upper estimates for the coderivative of the solution map (1.3) to the initial QVI (1.2) with $\Gamma(x,y)$ given in (3.1). The results obtained in what follows are largely based on the coderivative estimates for the multivalued term (3.4) of this QVI established in section 3 via coderivative calculus.

To begin, we consider the *parameter-dependent GE*

$$0\in g(x,y)+Q(x,y) \tag{4.1}$$

with both single-valued term $g\colon \mathbb{R}^n\times\mathbb{R}^m\to\mathbb{R}^m$ and set-valued term $Q\colon \mathbb{R}^n\times\mathbb{R}^m\rightrightarrows\mathbb{R}^m$ depending on the parameter $x\in\mathbb{R}^n$, where g is smooth as in section 1, while $Q\colon \mathbb{R}^n\times\mathbb{R}^m\rightrightarrows\mathbb{R}^m$ is an arbitrary set-valued mapping, which *may not* be in the special form (3.4). The following lemma provides a coderivative upper estimate for the solution map to the parameter-dependent GE (4.1) via the *adjoint Jacobian* of g and the *coderivative* of Q under the appropriate *calmness* assumption.

LEMMA 4.1 (coderivatives of solution maps to parameter-dependent GEs). *Let $(\bar{x},\bar{y})$ satisfy the GE (4.1), and let*

$$S(x):=\{y\in\mathbb{R}^m |\ 0\in g(x,y)+Q(x,y)\}$$

be the solution map to this GE. Assume that g is continuously differentiable around $(\bar{x},\bar{y})$, that Q is locally closed-graph around this point, and that the set-valued mapping $\Xi\colon \mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}^m\rightrightarrows\mathbb{R}^n\times\mathbb{R}^m$ defined by

$$\Xi(\vartheta):=\left\{(x,y)\in\mathbb{R}^n\times\mathbb{R}^m \;\middle|\; \begin{bmatrix} x \\ y \\ -g(x,y) \end{bmatrix}+\vartheta\in\operatorname{gph} Q\right\} \tag{4.2}$$

is calm at $(0,\bar{x},\bar{y})$. *Then for all* $u\in\mathbb{R}^m$ *one has the estimate*

$$D^*S(\bar{x},\bar{y})(u)\subset\Bigg\{v\in\mathbb{R}^n\;\Big|\;\exists\, w\in\mathbb{R}^m \text{ with } \begin{bmatrix} v\\ -u\end{bmatrix}\in(\nabla g(\bar{x},\bar{y}))^Tw + D^*Q(\bar{x},\bar{y},-g(\bar{x},\bar{y}))(w)\Bigg\}. \tag{4.3}$$

Proof. Under the imposed calmness assumption, it holds by [16, Theorem 4.1] that

$$N_{\Xi(0)}(\bar{x},\bar{y})\subset\begin{bmatrix}E & 0 & -(\nabla_x g(\bar{x},\bar{y}))^T\\ 0 & E & -(\nabla_y g(\bar{x},\bar{y}))^T\end{bmatrix}\circ N_{\operatorname{gph}Q}(\bar{x},\bar{y},-g(\bar{x},\bar{y})).$$

Since $\operatorname{gph}S=\Xi(0)$ for the solution map to the GE (4.1), we arrive at the coderivative upper estimate (4.3). □

Remark 4.2 (more on coderivatives of solution maps to GEs). As above, we can present efficient conditions ensuring the fulfillment of the *calmness* requirement of Lemma 4.1. First it is automatic under the *polyhedrality* assumptions (in finite dimensions). Furthermore, the stronger *Lipschitz-like* property holds by the coderivative criterion (2.7) applied to the mapping Ξ in (4.2) in both finite and infinite dimensions, even for *nonsmooth* mappings g with replacing the adjoint Jacobian $\nabla g(\bar{x},\bar{y})^Tw$ by the coderivative $D^*g(\bar{x},\bar{y})(w)$ of g or its scalarization $\partial\langle w,g\rangle(\bar{x},\bar{y})$. In the framework of (4.2), the latter criterion amounts to saying (similarly to the *Fredholm alternative*) that the *adjoint GE*

$$0\in\nabla g(\bar{x},\bar{y})^Tw+D^*Q(\bar{x},\bar{y},-g(\bar{x},\bar{y}))(w)$$

has *only the trivial solution* $w=0$; see [26, Theorem 4.46]. Observe also that there are certain conditions given in [26, Theorem 4.44 and Corollary 4.45] ensuring the *equality* in (4.3). Unfortunately, the efficient realization of these conditions for *parameter-dependent* mappings $Q=Q(x,y)$ in (4.1) requires the *graph-convexity* of Q, which is not often the case for mappings (3.4) arising in QVIs. For *parameter-independent* $Q=Q(y)$ in (4.1), the equality in (4.3) is achieved under the *surjectivity* (full rank) requirement of the Jacobian matrix $\nabla g(\bar{x},\bar{y})$.

Now we proceed by studying the *solution map* (1.3) to the *QVI* (1.1) generated by the parameterized set Γ from (3.1). To simplify formulas of type (4.3) in what follows, we omit the text corresponding to "$\exists w\in\mathbb{R}^m$ with." By the same reason we introduce the *Lagrangian* mapping $\mathcal{L}\colon\mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}^s\to\mathbb{R}^m$ defined by

$$\mathcal{L}(x,y,d):=g(x,y)+(p(x,y))^Td. \tag{4.4}$$

Therefore the adjoint Lagrangian partial derivative is represented as

$$(\nabla_{x,y}\mathcal{L}(x,y,d))^T=(\nabla g(x,y))^T+(\nabla_{x,y}((p(x,y))^Td))^T.$$

To formulate the main result of this section, define the mapping $\Lambda\colon\mathbb{R}^n\times\mathbb{R}^m\rightrightarrows\mathbb{R}^s$ by

$$\Lambda(x,y):=\{d\in\mathbb{R}^s|\;\mathcal{L}(x,y,d)=0\}. \tag{4.5}$$

THEOREM 4.3 (coderivative estimate for solution maps to QVIs). *Let* $S\colon\mathbb{R}^n\rightrightarrows\mathbb{R}^m$ *be the solution map* (1.3) *to the original QVI represented by* (3.3) *around the reference*

point $(\bar{x},\bar{y})\in\operatorname{gph}S$ *with the Lagrangian* $\mathcal{L}$ *defined in* (4.4). *Assume that the CQ* (3.6) *holds and that the multifunction* M *given by* (3.7) *is calm at all the points* $(0,\bar{x},\bar{y},d)$ *with* $d\in\Lambda(\bar{x},\bar{y})$. *We also suppose that the multifunction* $P\colon\mathbb{R}^m\times\mathbb{R}^s\times\mathbb{R}^s\rightrightarrows\mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}^s$ *defined by*

$$P(z,\vartheta):=\Big\{(x,y,d)\in\mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}^s\;\Big|\;\mathcal{L}(x,y,d)+z=0\Big\}\cap M(\vartheta)\tag{4.6}$$

is calm at the points $(0,0,\bar{x},\bar{y},d)$ *with* $d\in\Lambda(\bar{x},\bar{y})$. *Then for all* $u\in\mathbb{R}^m$ *we have*

$$\begin{aligned}D^*S(\bar{x},\bar{y})(u)\subset\bigcup_{d\in\Lambda(\bar{x},\bar{y})}\Big\{&(\nabla_x\mathcal{L}(\bar{x},\bar{y},d))^Tv+(\nabla_xr(\bar{x},\bar{y},d))^Tw\;\Big|\;0=u\\&+(\nabla_y\mathcal{L}(\bar{x},\bar{y},d))^Tv+(\nabla_yr(\bar{x},\bar{y},d))^Tw,\;\;w\in D^*N_\Theta(r(\bar{x},\bar{y}),d)(p(\bar{x},\bar{y})v)\Big\}.\end{aligned}\tag{4.7}$$

If, furthermore, the mappings $g(\cdot)$ *and* $r(\cdot)$ *are affine and the set* Θ *is polyhedral, then the above calmness requirements are automatic, and it suffices to assume only the CQ* (3.6).

Proof. First observe, similarly to the proof of Corollary 3.2 in case (b), that both calmness conditions imposed in the theorem are automatically fulfilled when the mappings g and r are affine and the set Θ is polyhedral. It remains to justify (4.7) under the calmness assumptions made.

We can easily see that inclusion (4.7) follows from assertion (ii) of Theorem 3.1 and Lemma 4.1 *provided that* the multifunction Ξ from (4.2) with Q given by (3.4) is *calm* at $(0,\bar{x},\bar{y})$. This can be ensured by requiring that the *expanded* multifunction $\overline{P}\colon\mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}^m\rightrightarrows\mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}^s$ defined by

$$\begin{aligned}\overline{P}(\vartheta_1,\vartheta_2,\vartheta_3):=\Big\{(x,y,d)\in\mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}^s\;\Big|\;&-g(x,y)+\vartheta_3\\&=(p(x+\vartheta_1,y+\vartheta_2))^Td,\;\;d\in N_\Theta(r(x+\vartheta_1,y+\vartheta_2))\Big\}\end{aligned}$$

is calm at all the points $(0,0,0,\bar{x},\bar{y},d)$ with $d\in\Lambda(\bar{x},\bar{y})$. Indeed, in the case under consideration we have the representation

$$\begin{aligned}\Xi(\vartheta_1,\vartheta_2,\vartheta_3)=\Big\{(x,y)\in\mathbb{R}^n\times\mathbb{R}^m\;\Big|\;&\exists\,d\in N_\Theta(r(x+\vartheta_1,y+\vartheta_2))\\&\text{with}\;\;-g(x,y)+\vartheta_3=(p(x+\vartheta_1,y+\vartheta_2))^Td\Big\},\end{aligned}$$

which implies the relationship

$$\Xi(\vartheta_1,\vartheta_2,\vartheta_3)=\operatorname{proj}_{x,y}\overline{P}(\vartheta_1,\vartheta_2,\vartheta_3).$$

Invoking now the result of [17, Lemma 1], we conclude that the required calmness property of the mapping $\overline{P}$ is implied by the calmness of the (only *canonically perturbed*) mapping $\widehat{P}\colon R^m\times\mathbb{R}^s\times\mathbb{R}^s\rightrightarrows\mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}^s$ defined by

$$\begin{aligned}&\widehat{P}(w_1,w_2,w_3)\\&\quad:=\left\{(x,y,d)\in\mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}^s\;\middle|\;\mathcal{L}(x,y,d)+w_1=0,\;\begin{bmatrix}r(x,y)+w_2\\d+w_3\end{bmatrix}\in\operatorname{gph}N_\Theta\right\}\end{aligned}$$

at the points $(0,0,0,\bar{x},\bar{y},d)$ with $d\in\Lambda(\bar{x},\bar{y})$. It remains to observe that

$$P(z,\vartheta)=\widehat{P}(w_1,w_2,w_3)$$

for $z = w_1$ and $\vartheta = (w_2, w_3)$, which completes the proof of the theorem. □

Similarly to the proof of Corollary 3.2 in case (a), we can get, on the basis of the coderivative criterion for the Lipschitz-like property (2.7), a *second-order* CQ ensuring the calmness property of the mapping P in (4.6). This will be done in the next section as a part of deriving a unified condition that implies both calmness properties assumed in Theorem 4.3 and simultaneously yields robust Lipschitzian stability of the solution map to the QVI under consideration.

It is clear that the above results can be used also in the case of

$$\Gamma(x) := \{y \in \mathbb{R}^m |\ q(x,y) \in \Theta\} \tag{4.8}$$

in (3.1) when (1.2) reduces to the parameter-dependent *variational inequality*

$$0 \in g(x,y) + N_{\Gamma(x)}(y), \quad y \in \Gamma(x),$$

which has been intensively investigated in many publications; see, e.g., [10] with the references therein and also [20] for the coderivative analysis of such variational inequalities and its application to parametric optimization. As a special case of such a variational inequality we can consider the *projection mapping* $\pi\colon \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^m$ defined by

$$\pi(z,x) := \mathrm{proj}_{\,\Gamma(x)}(z), \tag{4.9}$$

with $\Gamma(x)$ given in (4.8); see, e.g., [10, 20] for the importance of the projection mapping (4.9) and its coderivative in various applications. Observe that the projection mapping (4.9) happens to be *single-valued* due to the *convexity* of the sets $\Gamma(x)$. On the basis of Theorem 4.3 we now derive an upper estimate of the coderivative of π.

THEOREM 4.4 (coderivatives of projection mappings in variational inequalities). *Given $\bar{x} \in \mathrm{dom}\,\Gamma$ and $\bar{z} \in \mathbb{R}^m$, put $\bar{y} := \pi(\bar{z}, \bar{x})$ and impose the CQ*

$$\left.\begin{array}{c}(\nabla_y q(\bar{x},\bar{y}))^T u = 0\\ u \in N_\Theta(q(\bar{x},\bar{y}))\end{array}\right\} \Longrightarrow u = 0.$$

Define the mappings $r\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^s$, $p\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^{m+s}$, and $Q\colon \mathbb{R}^n \times \mathbb{R}^m \rightrightarrows \mathbb{R}^m$ by

$$r(x,y) := q(x,y), \quad p(x,y) := \nabla_y q(x,y), \quad Q(x,y) := (\nabla_y q(x,y))^T N_\Theta(q(x,y))$$

and assume that the multifunction $M(\vartheta)$ given by (3.7) via this mappings r is calm at every point $(\bar{x},\bar{y},d)$ with $d \in N_\Theta(r(\bar{x},\bar{y}))$ and $(p(\bar{x},\bar{y}))^T d = \bar{z} - \bar{y}$. Then for all $u \in \mathbb{R}^m$ we have the coderivative upper estimate

$$D^*\pi(\bar{z},\bar{x},\bar{y},)(u) \subset \bigcup_{\substack{d \in N_\Theta(q(\bar{x},\bar{y}))\\ (\nabla_y q(\bar{x},\bar{y}))^T d = \bar{z}-\bar{y}}} \left\{ (t,v) \in \mathbb{R}^m \times \mathbb{R}^n \;\middle|\; t = -w, \right.$$
$$\left. \begin{bmatrix} v \\ -u - w \end{bmatrix} \in D^*Q(\bar{x},\bar{y},\bar{z}-\bar{y})(w) \right\}, \tag{4.10}$$

where the coderivative D^*Q is in turn estimated by (3.8).

Proof. It is easy to observe from the construction of (4.9) and the normal cone definition in convex analysis that the projection relation $y = \pi(z, x)$ can be equivalently written as the *parameterized GE*

$$0 \in y - z + N_{\Gamma(x)}(y), \tag{4.11}$$

where besides x there is an additional parameter z. Thus we can apply the results of Lemma 4.1 and Theorem 4.3 to estimate the coderivative of the solution map to (4.11). To proceed, we need to check the two calmness requirements of Theorem 4.3 in the case of (4.11). Note that the presence of the additional parameter z has no influence on the first calmness requirement. The second one reduces for (4.11) to the calmness of the multifunction $P\colon \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m \rightrightarrows \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^m$ defined by

$$P(\vartheta) := \left\{ (z, x, y) \in \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^m \;\middle|\; \begin{bmatrix} x \\ y \\ z - y \end{bmatrix} + \vartheta \in \operatorname{gph} Q \right\} \tag{4.12}$$

at the point $(0, \bar{z}, \bar{x}, \bar{y})$. The latter holds, however, automatically due to the inclusion

$$D^*P(0, \bar{z}, \bar{x}, \bar{y})(0) \subset \left\{ v = (v_1, v_2, v_3) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m \;\middle|\; 0 = \begin{bmatrix} 0 & 0 & E \\ E & 0 & 0 \\ 0 & E & -E \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}, \right.$$
$$\left. v \in N_{\operatorname{gph} Q}(\bar{x}, \bar{y}, \bar{z} - \bar{y}) \right\} = \{0\},$$

which ensures the Lipschitz-like property of the mapping (4.12) around $(0, \bar{z}, \bar{x}, \bar{y})$ by the coderivative criterion (2.7). The coderivative estimate (4.10) follows now from the modified inclusion (4.3) in the case of (4.11). □

5. Robust Lipschitzian stability of QVIs. By *robust Lipschitzian stability* of QVIs we understand in this paper the fulfillment of the Lipschitz-like property of the solution map (1.3) to the QVI (1.2) with the generating sets $\Gamma(x, y)$ given by (3.1) around the reference point $(\bar{x}, \bar{y})$. This type of Lipschitzian behavior has been well recognized as an appropriate stability property of local sensitivity analysis, which is *robust* (i.e., *preserved*) under small parameter perturbations. As mentioned in section 2, the Lipschitz-like property can be viewed as a graph localization of the classical *local Lipschitzian* behavior being closely related to the two other major *well-posedness* properties of nonlinear analysis known as *metric regularity* and *linear openness*; see [26, 37] for more details, discussions, and references.

To derive efficient conditions for robust Lipschitzian stability of the QVIs under consideration, we utilize in what follows the *coderivative criterion* (2.7) for the Lipschitz-like property combined with the constructive *coderivative estimate* for the solution map (1.3) established in the previous section on the base of new rules of coderivative calculus. Furthermore, the coderivative results developed above allow us to conduct not only qualitative but also *quantitative* analysis of robust Lipschitzian stability for QVIs by providing an estimate of the *exact Lipschitzian bound*. The second assertion of the following theorem contains efficient conditions, which simultaneously ensure the desired Lipschitzian stability and the calmness requirements needed for the validity of the crucial coderivative estimate established in Theorem 4.3.

THEOREM 5.1 (Lipschitzian stability of solution maps). *Let S be the solution map* (1.3) *to the QVI under consideration, with Γ given by* (3.1) *and with $(\bar{x}, \bar{y}) \in \operatorname{gph} S$. The following assertions hold:*

(i) *Suppose that all the assumptions of Theorem* 4.3 *are satisfied. Then the solution map S is Lipschitz-like around $(\bar{x},\bar{y})$ provided that*

$$\left.\begin{aligned} &0=(\nabla_y\mathcal{L}(\bar{x},\bar{y},d))^Tv+(\nabla_y r(\bar{x},\bar{y}))^Tw\\ &d\in\Lambda(\bar{x},\bar{y})\\ &w\in D^*N_\Theta(r(\bar{x},\bar{y}),d)(p(\bar{x},\bar{y})v)\end{aligned}\right\}\Longrightarrow(\nabla_x\mathcal{L}(\bar{x},\bar{y},d))^Tv+(\nabla_x r(\bar{x},\bar{y}))^Tw=0.$$

Furthermore, we have the upper estimate

$$\begin{aligned}\operatorname{lip}S(\bar{x},\bar{y})\le\sup\Big\{&\|(\nabla_x\mathcal{L}(\bar{x},\bar{y},d))^Tv+(\nabla_x r(\bar{x},\bar{y},d))^Tw\|\ \Big|\\ &d\in\Lambda(\bar{x},\bar{y}),\ \ 0\in u+(\nabla_y\mathcal{L}(\bar{x},\bar{y},d))^Tv+(\nabla_y r(\bar{x},\bar{y}))^Tw,\\ &w\in D^*N_\Theta(r(\bar{x},\bar{y}),d)(p(\bar{x},\bar{y})v),\ \ \|u\|\le1\Big\}\end{aligned}\tag{5.1}$$

for the exact Lipschitzian bound (2.8) *of the solution map in* (1.3), (3.1).

(ii) *Suppose that the CQ* (3.6) *is satisfied and that*

$$\left.\begin{aligned} &0=(\nabla_y\mathcal{L}(\bar{x},\bar{y},d))^Tv+(\nabla_y r(\bar{x},\bar{y}))^Tw\\ &d\in\Lambda(\bar{x},\bar{y})\\ &w\in D^*N_\Theta(r(\bar{x},\bar{y}),d)(p(\bar{x},\bar{y})v)\end{aligned}\right\}\Longrightarrow v=0,\ w=0.\tag{5.2}$$

Then the solution map S is Lipschitz-like around $(\bar{x},\bar{y})$ with the bound estimate (5.1).

Proof. To justify both statements in (i), recall that under the assumptions of Theorem 4.3 we have the coderivative estimate (4.7). Substituting the set on the right-hand side of (4.7) into the coderivative criterion (2.7) and the exact bound formula (2.8), we arrive at the sufficient condition for the Lipschitz-like property in (i) and the bound estimate (5.1).

To justify (ii), we first apply the coderivative criterion (2.7) to the mapping P in (4.6) and observe that the Lipschitz-like property of P around $(0,0,\bar{x},\bar{y},d), d\in\Lambda(\bar{x},\bar{y})$, and hence its calmness at this point required by Theorem 4.3, is ensured by the implication

$$\left.\begin{aligned} &0=(\nabla_{x,y}\mathcal{L}(\bar{x},\bar{y},d))^Tv+(\nabla r(\bar{x},\bar{y}))^Tw\\ &w\in D^*N_\Theta(r(\bar{x},\bar{y}),d)(p(\bar{x},\bar{y})v)\end{aligned}\right\}\Longrightarrow v=0,w=0.\tag{5.3}$$

This condition ensures simultaneously also the CQ (3.12) with $(p(\bar{x},\bar{y}))^Td=-g(\bar{x},\bar{y})$. It is now easy to conclude that implication (5.2) together with the CQ (3.6) imply the fulfillment of all requirements posed in assertion (i) of this theorem. Thus we arrive at both statements in (i) under (3.6) and (5.2) and complete the proof of the theorem. □

Let us illustrate the application of Theorem 5.1 to the study of robust Lipschitzian stability of the solution map to a QVI arising in *Nash equilibrium* modeling.

Example 5.2 (Lipschitzian stability in parameterized Nash games). Consider the parameter-dependent QVI of type (1.2) given as

$$0\in\begin{bmatrix}-34+2y^1+\frac{8}{3}y^2\\ -\frac{97}{4}+\frac{5}{4}y^1+2y^2\end{bmatrix}+N_{\Gamma(x,y)}(y),\quad x\in\mathbb{R},\ y=(y^1,y^2)\in\mathbb{R}^2,\tag{5.4}$$

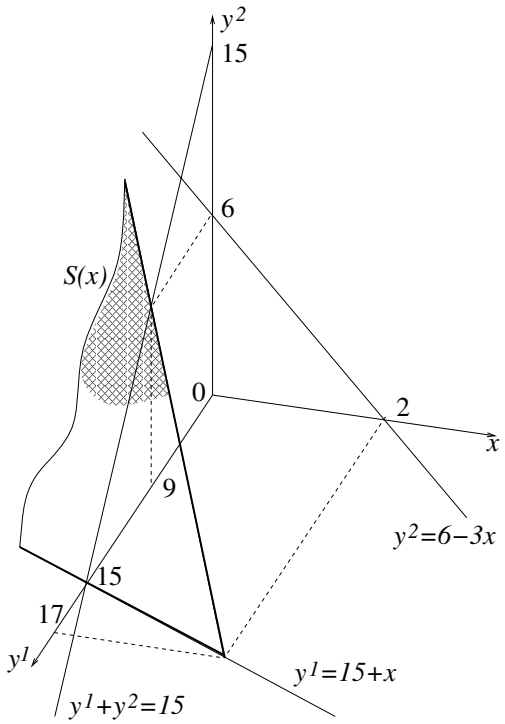


FIG. 1. *Local behavior of* $\operatorname{gph} S$ *around* $(0,9,6)$.

with $\Gamma(x,y)=\Gamma(x,y^1,y^2)$ defined by

$$\Gamma(x,y^1,y^2):=\big\{(z^1,z^2)\in\mathbb{R}^2\big|\ y^1+z^2-15-x\le 0,\ y^2+z^1-15-x\le 0\big\}.$$

This QVI is taken from [14] and corresponds to a *parameterized Nash game of two players* with the parameter-independent objectives

$$(y^1)^2-34y^1+\frac{8}{3}y^2,\quad \frac{5}{4}y^1-\frac{97}{4}y^2+(y^2)^2$$

subject to the parameter-dependent inequality constraint

$$y^1+y^2\le 15+x;$$

see [14] for more details and economic descriptions. Observe that the corresponding set-valued mapping $Q(x,y)$ from (3.4) is fully considered in Example 3.3 with constructively estimating its coderivative by Theorem 3.1.

Take the reference point $(\bar{x},\bar{y})=(0,9,6)$ (which is actually a local *optimal solution* to an MPEC involving the QVI constraint (5.4); see Example 6.3 for more details) and investigate the robust Lipschitzian stability of the solution map S to the QVI (5.4) around this point by employing the results of Theorem 5.1. S is depicted in Figure 1. We have $g(\bar{x},\bar{y})=\left[\begin{smallmatrix}0\\-1\end{smallmatrix}\right]$ for the single-valued term of (5.4), which is affine together with the function $q(x,y,z)$ describing the generating sets $\Gamma(x,y)$ in (5.4) via the polyhedral set $\Theta=\mathbb{R}^2{}_-$. By Theorem 5.1, the Lipschitz-like property of the solution map to (5.4) around the reference point $(\bar{x},\bar{y})$ is ensured by the implication

in assertion (i). It reduces to

$$\left.\begin{aligned} 0 &= 2v^1 + \tfrac{5}{4}v^2 + w \\ 0 &= \tfrac{8}{3}v^1 + 2v^2 + w \\ 0 &= v^2 \end{aligned}\right\} \Longrightarrow -w = 0,$$

which is trivially satisfied, and thus we have the desired stability.

Remark 5.3 (extensions and modifications of stability results for QVIs). Let us discuss some directions of possible extensions and modifications of the Lipschitzian stability results for QVIs obtained in this section.

(i) Based on the coderivative representations/estimates for solution maps to generalized equations of type (4.1) derived in [26, subsection 4.4.1], on the new results of this paper for coderivative estimates of the multivalued term in the QVI (1.2), and on the coderivative characterization (2.7) and (2.8) of the Lipschitz-like property, we can extend the Lipschitzian stability results obtained above to the case of *nonsmooth* mappings g in (1.2). Results in this direction involve the coderivative of g (or its subdifferential scalarization) replacing the adjoint Jacobian.

(ii) Another approach to robust Lipschitzian stability, generally *independent* of the one used above, can be developed in the framework of QVIs under consideration. This approach, which is the most efficient in the case of *canonical perturbations*, involves a preliminary *strong approximation* (in Robinson's sense [36]) of the single-valued term in the GE (1.2) and then employs the coderivative calculus for the multivalued term of (1.2) established in section 3. We refer the reader to [25] and [26, subsection 4.4.3] for more details in the case of GEs (4.1) and some of their specifications and emphasize that the efficient implementation of this approach (as well as the one developed in this paper) is based on the new coderivative calculus results for the multivalued term in (1.2) obtained in section 3.

(iii) The reader can observe that the *calmness* property of the corresponding mappings plays a significant role in the main results of the paper. The efficient realization of this property comes into play via either a kind of the *generalized Mangasarian–Fromovitz CQ* or via *polyhedrality*. They are indeed the two major cases of calmness; some other examples and related discussions can be found in [16, 17, 32]. It should be noted in this respect that the calmness property definitely requires further investigation, which should address first of all *calculus* issues for this property ensuring its preservation under various operations performed on set-valued mappings.

6. Optimality conditions for mathematical programs with QVI constraints. In this section we study a class of parametric optimization problems of the special type

$$\begin{cases} \text{minimize } \ \varphi(x,y) \\ \text{subject to} \\ 0 \in g(x,y) + N_{\Gamma(x,y)}(y), \quad (x,y) \in \Omega, \end{cases} \tag{6.1}$$

where $\varphi\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is an *objective/cost function*, and where $\Omega \subset \mathbb{R}^n \times \mathbb{R}^m$ is an nonempty, closed set imposing the so-called *geometric constraints* in (6.1). The moving set $\Gamma(x,y)$ under the normal cone operation in (6.1) is taken from (3.1). It generates the most crucial *QVI constraints* in (6.1), which incorporate the optimization problem (6.1) into a broad class of MPECs, whose importance has been well recognized in optimization theory and applications; see, e.g., the books [10, 21, 27, 33] with many references, examples, and discussions therein.

The underlying feature if the QVI constraints in (6.1) is that they are given in the form of *parametric GEs*

$$0 \in g(x,y) + Q(x,y) \;\text{ with }\; (x,y) \in \Omega, \tag{6.2}$$

where not only the single-valued term g but also the *set-valued term* Q *depend on the parameter* $x \in \mathbb{R}^n$. The latter class of MPECs has not attracted much attention in the literature; see [27] for quite recent results and commentaries. Furthermore, the results of [27, section 5.2] concerning necessary optimality conditions for MPECs governed by parameter-dependent generalized equations of type (6.2) cannot be directly applied to the QVI problem (6.1) without *new coderivative calculus rules* taking into account the specific nature of mappings $Q(x,y) = N_{\Gamma(x,y)}(y)$ under consideration, which are in fact the *main thrust* of this paper.

In what follows we pay the main attention to deriving *necessary optimality conditions* for local minimizers to problem (6.1), assuming, for brevity and simplicity, that the cost function φ is *continuously differentiable* around the reference point. There is plenty of room for extending the results obtained below to the case of nonsmooth cost functions φ (as well as nonsmooth mappings g in the QVI constraints of this problem); see Remark 6.2(iii) for more discussions.

In the following theorem we use the notation for the QVI data in the constraints of (6.1) introduced in sections 3 and 4.

THEOREM 6.1 (necessary conditions for optimal solutions to MPECs with QVI constraints). *Let* $(\bar{x},\bar{y})$ *be a local optimal solution to problem* (6.1). *Assume that the partial Jacobian matrix* $\nabla_3 g(\bar{x},\bar{y},\bar{y})$ *has the maximal/full row rank, and let* $\bar{d} \in \mathbb{R}^s$ *be the unique vector satisfying the equation*

$$\mathcal{L}(\bar{x},\bar{y},d) = 0, \tag{6.3}$$

with the Lagrangian $\mathcal{L}$ *defined in* (4.4). *Suppose further that the multifunction* M *from* (3.7) *is calm at* $(0,\bar{x},\bar{y},\bar{d})$ *and that the multifunction* $\widetilde{P}$ *defined by*

$$\widetilde{P}(z,\vartheta) := \Big\{(x,y,d) \in \Omega \times \mathbb{R}^s \;\Big|\; \mathcal{L}(x,y,d) + z = 0\Big\} \cap M(\vartheta)$$

is calm at $(0,0,\bar{x},\bar{y},\bar{d})$. *Then there exist multipliers* $v \in \mathbb{R}^m$ *and* $u \in \mathbb{R}^s$ *such that*

$$\begin{aligned} 0 &\in \nabla\varphi(\bar{x},\bar{y}) + (\nabla\mathcal{L}(\bar{x},\bar{y},\bar{d}))^T v + (\nabla r(\bar{x},\bar{y}))^T u + N_\Omega(\bar{x},\bar{y}),\\ 0 &\in -u + D^* N_\Theta(r(\bar{x},\bar{y}),\bar{d})(p(\bar{x},\bar{y})v). \end{aligned} \tag{6.4}$$

If, moreover, the mappings g, r *are affine and the sets* Θ, Ω *are polyhedral, then both calmness assumptions above hold automatically.*

Proof. Let Σ be the set of feasible solutions to (6.1) given by

$$\Sigma := \left\{(x,y) \in \Omega \;\middle|\; \begin{bmatrix} x\\ y\\ -g(x,y)\end{bmatrix} \in \operatorname{gph} Q\right\} \;\text{ with }\; Q(x,y) = N_{\Gamma(x,y)})(y). \tag{6.5}$$

We obviously have $\Sigma = \Omega \cap \Xi(0)$, where the mapping $\Xi(\vartheta)$ is defined by (4.2). Following the proof of Theorem 4.3 (based on Lemma 4.1 and Theorem 3.1) and employing [16, Theorem 4.1], which ensures the required calculus rules under the imposed *calmness* assumptions, we arrive at the normal cone upper estimate

$$\begin{aligned} N_\Sigma(\bar{x},\bar{y}) \subset \Big\{&(\nabla\mathcal{L}(\bar{x},\bar{y},\bar{d}))^T v + (\nabla r(\bar{x},\bar{y}))^T u \;\Big|\\ &u \in D^* N_\Theta(r(\bar{x},\bar{y}),\bar{d})(p(\bar{x},\bar{y})v)\Big\} + N_\Omega(\bar{x},\bar{y}). \end{aligned} \tag{6.6}$$

Observe that the MPEC (6.1) can be equivalently written as the optimization problem

$$\text{minimize } \varphi(x,y) \text{ subject to } (x,y)\in\Sigma, \tag{6.7}$$

with only the *geometric constraint* defined by (6.5). By [27, Proposition 5.1] one has the necessary optimality condition

$$0\in\nabla\varphi(\bar{x},\bar{y})+N_\Sigma(\bar{x},\bar{y}) \tag{6.8}$$

for the local solution $(\bar{x},\bar{y})$ to the latter (and the original) problem. Substituting (6.6) into (6.8), we arrive at (6.4). The fulfillment of the imposed calmness requirements under the linearity/polyhedrality assumptions in the last statement of the theorem is justified similarly to the proof of Corollary 3.2 in case (b). □

Remark 6.2 (discussions on optimality conditions for problems with QVI constraints). Let us discuss some specific features and possible extensions of the necessary conditions for MPECs with QVI constraints obtained in Theorem 6.1.

(i) By now it has been well recognized that the *calmness* assumptions play an important role as *CQs* for MPECs. Necessary conditions of type (6.4) with a calmness CQ were derived for the first time in [38] in the case of MPECs with variational inequality constraints. Observe that if Θ is a cone with vertex at the origin, the relation $d\in N_\Theta(r(x,y))$ can be replaced by $r(x,y)\in N_{\Theta^*}(d)$ via the *conjugate/dual* (or negative polar) cone Θ^* to Θ. This enables us to derive, following the approach in [32], the optimality conditions in the above theorem, which hold without the first calmness assumption in this case. We refer the reader to [27, subsection 5.2.3] for recent necessary optimality conditions in general MPECs with parameter-dependent constraints (6.2) under calmness CQs whose implementation in the case of QVI constraints requires the coderivative calculus rules developed in this paper.

(ii) The *full row rank* requirement imposed in Theorem 6.1 obviously ensures the fulfillment of the CQ (3.6) and the unique solvability of the Lagrangian equation (6.3) for $d\in\mathbb{R}^s$. If $\Gamma=\Gamma(x)$ and $\Theta=\mathbb{R}^s_-$, the latter reduces to the classical *strict Mangasarian–Fromovitz CQ*. In principal, we can proceed in the proof of Theorem 6.1 without the above full rank assumption for general set-valued mappings Q; see, e.g., [26, subsection 3.1.1] for computing/estimating the normal cone (2.1) to arbitrary closed sets of type (6.5). It happens, however, that the *upper estimate* (3.8) of the coderivative of Q arising in the QVI constraint of (6.1) may be *very poor* in the absence of the full rank condition. This phenomenon can be observed not only in the case of QVIs but also for standard variational inequalities with $\Gamma(x)$ given by (4.8) with $n=m=1$, $s=2$, $\Theta=\mathbb{R}^2_-$, and

$$q(x,y)=\begin{bmatrix} y-x\\ y+x-2\end{bmatrix}.$$

One can check that in this case the right-hand side of the coderivative estimate (3.8) for Q is *all* $\mathbb{R}$, although the required assumptions (except the full rank one) are satisfied; cf. [8].

(iii) Following the MPEC developments in [27, section 5.2] and having in hand the new calculus results obtained in this paper, we *do not have any difficulties* extending the necessary optimality conditions of Theorem 6.1 for QVI optimization problems of type (6.1) to the case of *nonsmooth cost functions* φ (as well as to the case of nonsmooth constraint functions g in (6.1); see the discussions in Remark 4.2 for the latter case). Indeed, the proof of Theorem 6.1 reduces the matter to deriving necessary

optimality conditions for the optimization problem (6.7) with merely *geometric constraints* of the *special type* handled by the calculus rules of this paper. Concerning the geometric constrained problem (6.7) per se, *two major types* of necessary optimality conditions are derived in [27] for this problem in general nonsmooth frameworks: the so-called *lower subdifferential* and *upper subdifferential* optimality conditions. These results allow us to replace (6.8) by its far-reaching extensions to nonsmooth cost functions, which are formulated via lower and upper subgradients of φ at the reference local minimizer; see [27, sections 5.1 and 5.2] for precise formulations and discussions. In this way we can readily extend Theorem 6.1 to nonsmooth problems with QVI constraints.

We conclude this paper by applications of Theorem 6.1 to two optimization models with parameter-dependent QVI constraints arising in practical situations. The first model concerns an MPEC with QVI constraints describing the *Nash game of two players* [14] considered in Example 5.2 from the viewpoint of robust Lipschitzian stability.

Example 6.3 (optimality conditions for MPEC with Nash equilibrium constraints). Consider the MPEC (6.1) with the cost function

$$\varphi(x,y) := x - 3y^1 - \frac{11}{3}y^2 + \frac{1}{2}(y^1-9)^2, \quad x \in \mathbb{R}, \;\; y=(y^1,y^2) \in \mathbb{R}^2,$$

the equilibrium constraint governed by the QVI (5.4) from Example 5.2, and the geometric constraint $(x,y) \in [-1,1] \times \mathbb{R}^2$. The *numerical approach* developed in [12] and the corresponding *SQP code* SNOPT allow us to compute the local optimal solution $(\bar{x},\bar{y}) = (0,9,6)$ to the MPEC under consideration. Since all the assumptions of Theorem 6.1 are clearly fulfilled, we can verify the application of the necessary optimality conditions (6.4) in this setting. We have

$$\mathcal{L}(x,y,d) = \begin{bmatrix} -34 + 2y^1 + \frac{8}{3}y^2 + d^2 \\ -\frac{97}{4} + \frac{5}{4}y^1 + 2y^2 + d^1 \end{bmatrix}, \quad \nabla_x \mathcal{L}(x,y,d) = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

$$(\nabla_y \mathcal{L}(x,y,d))^T = \begin{bmatrix} 2 & \frac{5}{4} \\ \frac{8}{3} & 2 \end{bmatrix}, \quad (\nabla_x r(x,y))^T = [-1,-1], \quad (\nabla_y r(x,y))^T = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

Taking into account that $N_\Omega(\bar{x},\bar{y}) = \{0\}$ due to $\Omega = [-1,1] \times \mathbb{R}^2$ and the choice of $(\bar{x},\bar{y})$, we easily compute from the first inclusion in (6.4) that

$$v^1 = 1, \quad v^2 = 0, \quad u^1 + u^2 = 1.$$

The second line of (6.4) gives the relationships

$$0 \in -u^1 + D^* N_{\mathbb{R}_-}(0,1)v^2,$$
$$0 \in -u^2 + D^* N_{\mathbb{R}_-}(0,0)v^1,$$

and so we can put $u^1 = 1$ and $u^2 = 0$. Thus the reference point $(\bar{x},\bar{y})$ satisfies the necessary optimality conditions of Theorem 6.1.

The final example concerns an MPEC with QVI constraints related to *oligopolistic market equilibrium*; cf. [11, 30] for more detailed descriptions of this and related models.

Example 6.4 (optimization model for determining oligopolistic market equilibrium). Consider two firms sharing the same resource of input commodity (e.g., row

material) and an authority determining the amount of this commodity x available for the next time period. Let x_0 be a certain reasonable consumption of x, and let (y_0^1, y_0^2) be target productions of the firms announced in advance. The authority, playing the role of the *Leader*, may look for a reasonable trade-off between the *consumption excess* $\max\{0, x - x_0\}$ and the *production differences* $y^i - y_0^i,\ i = 1, 2$, where y^i is the actual production of the ith firm. The firms, being *Followers* in this game, behave *noncooperatively* and compute their productions y^1, y^2 by solving the QVI

$$0 = \begin{bmatrix} k_1 - (a - b(y^1 + y^2)) + y^1 b \\ k_2 - (a - b(y^1 + y^2)) + y^2 b \end{bmatrix} + N_{\Gamma(x,y)}(y), \tag{6.9}$$

where the positive numbers k_i specify the linear *production costs* of the firms, and where $a - b(y^1 + y^2)$ is the *inverse demand curve* that assigns to the quantity $y^1 + y^2$ available on the market the corresponding *price* at which consumers are willing to demand [30].

Let us specify the initial data of (6.9) by $k_1 = 24,\ k_2 = 28,\ a = 100,\ b = 1$, and

$$\Gamma(x, y) := \{(y^1, y^2) \in \mathbb{R}^2_+ \,|\, y^1 + y^2 \le 0.333x\},$$

and let the Leader's objective be given by

$$\varphi(x, y^1, y^2) := [\max\{0, x - x_0\}]^2 + \gamma[(y^1 - y_0^1)^2 + (y^2 - y_0^2)^2],$$

with $x_0 = 135,\ \gamma = 0.6,\ y_0^1 = 34$, and $y_0^2 = 16$.

Thus we arrive at the optimization problem of MPEC type (6.1) without any nonequilibrium (geometric) constraints. Using the *numerical approach* and *SQP code* mentioned in Example 6.3, compute the *optimal solution*

$$\bar{x} = 135.15, \quad \bar{y}^1 = 30.95, \quad \bar{y}^2 = 14.1$$

to the MPEC under consideration. Now applying Theorem 6.1, we have

$$r(x, y) = \begin{bmatrix} y^1 + y^2 - 0.333x \\ y^1 + y^2 - 0.333x \end{bmatrix}, \quad p(x, y) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

$$\mathcal{L}(x, y, d) = \begin{bmatrix} -76 + 2y^1 + y^2 \\ -72 + y^1 + 2y^2 \end{bmatrix} + \begin{bmatrix} d^2 \\ d^1 \end{bmatrix},$$

and then compute $\bar{d} = (12.85,\ 0)$ from (6.3). All the assumptions of Theorem 6.1 are clearly satisfied, and we can check that both relationships in (6.4) are fulfilled with $v = (1.38,\ 0)$ and $u = (0.3,\ 0.6)$. Indeed, the first relationship in (6.4) reduced to the system of linear equations

$$\begin{aligned} 0.333(-u^1 - u^2) &= -2(x - x_0), \\ 2v^1 + v^2 + u^1 + u^2 &= -2\gamma(y^1 - y_0^1), \\ v^1 + 2v^2 + u^1 + u^2 &= -2\gamma(y^2 - y_0^2), \end{aligned} \tag{6.10}$$

while the second one gives

$$u^1 \in D^* N_{\mathbb{R}_-}(0, 12.86)(v^2), \quad u^2 \in D^* N_{\mathbb{R}_-}(0, 0)(v^1). \tag{6.11}$$

From the first relationship in (6.11) we immediately get that $v^2 = 0$ and u^1 is free. The multipliers v^1, u^1, u^2 can be computed from the linear system (6.10). Since they fulfill the second relationship in (6.11), the optimality conditions have been verified.

Similarly we can consider oligopolistic market models with more realistic production costs and inverse demand curves taken, e.g., from [30]. This, however, requires more work for computing and verifications.

REFERENCES

[1] J.-P. Aubin, *Lipschitz behavior of solutions to convex minimization problems*, Math. Oper. Res., 9 (1984), pp. 87–111.

[2] C. Baiocchi and A. Capelo, *Variational and Quasivariational Inequalities. Applications to Free Boundary Problems*, J. Wiley & Sons, New York, 1984.

[3] A. Bensoussan and J.-L. Lions, *Nouvelle formulation des problemes de controle impulsionnel et applications*, C. R. Acad. Sci. Paris Sér. A-B, 276 (1973), pp. 1189–1192.

[4] P. Beremlijski, J. Haslinger, M. Kočvara, and J. Outrata, *Shape optimization in contact problems with Coulomb friction*, SIAM J. Optim., 13 (2002), pp. 561–587.

[5] D. Chan and J.-S. Pang, *The generalized quasivariational inequality problem*, Math. Oper. Res., 1 (1982), pp. 211–222.

[6] F. H. Clarke, *Optimization and Nonsmooth Analysis*, J. Wiley & Sons, New York, 1983.

[7] M. De Luca and A. Maugeri, *Discontinuous quasi-variational inequalities and applications to equilibrium problems*, in Nonsmooth Optimization: Methods and Applications, Gordon and Breach, Montreux, 1992, pp. 70–75.

[8] S. Dempe, J. Dutta, and S. Lohse, *Optimality Conditions for Bilevel Programming Programs*, Technical report, Department of Mathematics and Computer Science, TU Bergakademie, Freiberg, Germany, 2005.

[9] A. L. Dontchev and R. T. Rockafellar, *Characterizations of strong regularity for variational inequalities over polyhedral convex sets*, SIAM J. Optim., 6 (1996), pp. 1087–1105.

[10] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementary Problems*, Springer, New York, 2003.

[11] S. D. Flåm and B. Kummer, *Great Fish Wars and Nash Equilibria*, Technical report WP-0892, Department of Economics, University of Bergen, Norway, 1992.

[12] R. Fletcher and S. Leyffer, *Solving mathematical programs with complementarity constraints as nonlinear programs*, Optim. Methods Softw., 19 (2004), pp. 15–40.

[13] P. Hammerstein and R. Selten, *Game theory and evolutionary biology*, in Handbook of Game Theory with Economic Applications, Vol. 2, R. J. Aumann and S. Hart, eds., North–Holland, Amsterdam, 1994, pp. 929–993.

[14] P. T. Harker, *Generalized Nash games and quasivariational inequality*, European J. Oper. Res., 54 (1990), pp. 81–94.

[15] J. Haslinger and P. D. Panagiotopoulos, *The reciprocal variational approach to the Signorini problem with friction. Approximation results*, Proc. Roy. Soc. Edinburgh Sect. A, 98 (1984), pp. 365–383.

[16] R. Henrion, A. Jourani, and J. Outrata, *On the calmness of a class of multifunctions*, SIAM J. Optim., 13 (2002), pp. 603–618.

[17] R. Henrion and J. V. Outrata, *Calmness of constraint systems with applications*, Math. Program., 104 (2005), pp. 437–464.

[18] M. Kočvara and J. V. Outrata, *On optimization of systems governed by implicit complementarity problems*, Numer. Funct. Anal. Optim., 15 (1994), pp. 869–887.

[19] M. Kočvara and J. V. Outrata, *On a class of quasivariational inequalities*, Optim. Methods Softw., 5 (1995), pp. 275–295.

[20] A. B. Levy and B. S. Mordukhovich, *Coderivatives in parametric optimization*, Math. Program., 99 (2004), pp. 311–327.

[21] Z. Q. Luo, J.-S. Pang, and D. Ralph, *Mathematical Programs with Equilibrium Constraints*, Cambridge University Press, Cambridge, UK, 1996.

[22] B. S. Mordukhovich, *Approximation Methods in Problems of Optimization and Control*, Nauka, Moscow, 1988.

[23] B. S. Mordukhovich, *Complete characterizations of openness, metric regularity, and Lipschitzian properties of multifunctions*, Trans. Amer. Math. Soc., 340 (1993), pp. 1–35.

[24] B. S. Mordukhovich, *Generalized differential calculus for nonsmooth and set-valued mappings*, J. Math. Anal. Appl., 183 (1994), pp. 250–288.

[25] B. S. Mordukhovich, *Stability theory for parametric generalized equations and variational inequalities via nonsmooth analysis*, Trans. Amer. Math. Soc., 343 (1994), pp. 609–658.
[26] B. S. Mordukhovich, *Variational Analysis and Generalized Differentiation*, I: *Basic Theory*, Springer, Berlin, 2006.
[27] B. S. Mordukhovich, *Variational Analysis and Generalized Differentiation*, II: *Applications*, Springer, Berlin, 2006.
[28] B. S. Mordukhovich and J. V. Outrata, *On second-order subdifferentials and their applications*, SIAM J. Optim., 12 (2001), pp. 139–169.
[29] U. Mosco, *Implicit Variational Problems and Quasivariational Inequalities*, Springer, Berlin, 1976.
[30] F. H. Murphy, H. D. Sherali, and A. L. Soyster, *A mathematical programming approach for determining oligopolistic market equilibrium*, Math. Program., 24 (1982), pp. 92–106.
[31] J. V. Outrata, *Solution behaviour for parameter-dependent quasivariational inequalities*, RAIRO Rech. Opér., 30 (1996), pp. 399–415.
[32] J. V. Outrata, *A generalized mathematical program with equilibrium constraints*, SIAM J. Control Optim., 38 (2000), pp. 1623–1638.
[33] J. V. Outrata, M. Kočvara, and J. Zowe, *Nonsmooth Approach to Optimization Problems with Equilibrium Constraints*, Kluwer, Dordrecht, The Netherlands, 1998.
[34] S. M. Robinson, *Generalized equations and their solutions. Part* I: *Basic theory*, Math. Programming Stud., 10 (1979), pp. 128–141.
[35] S. M. Robinson, *Some continuity properties of polyhedral multifunctions*, Math. Programming Stud., 14 (1981), pp. 206–214.
[36] S. M. Robinson, *An implicit function theorem for a class of nonsmooth functions*, Math. Oper. Res., 16 (1991), pp. 292–309.
[37] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer, Berlin, 1998.
[38] J. J. Ye and X. Y. Ye, *Necessary optimality conditions for optimization problems with variational inequality constraints*, Math. Oper. Res., 22 (1997), pp. 977–997.

SIAM J. OPTIM.
Vol. 18, No. 2, pp. 413–436

ABSTRACT CONVEXITY AND AUGMENTED LAGRANGIANS*

REGINA SANDRA BURACHIK† AND ALEX RUBINOV‡

Abstract. The ultimate goal of this paper is to demonstrate that abstract convexity provides a natural language and a suitable framework for the examination of zero duality gap properties and exact multipliers of augmented Lagrangians. We study augmented Lagrangians in a very general setting and formulate the main definitions and facts describing the augmented Lagrangian theory in terms of abstract convexity tools. We illustrate our duality scheme with an application to stochastic semi-infinite optimization.

Key words. abstract convexity, nonconvex programming, Lagrange-type functions, augmented Lagrangians, exact penalty representation, stochastic semi-infinite programming

AMS subject classifications. 90C26, 52A01, 49N15, 65K10, 90C15

DOI. 10.1137/050647621

1. Introduction.

The classical augmented Lagrangian is defined by means of two addends; one of them is a classical Lagrangian, the other is a penalty term. Due to the term "augmented Lagrangian" the first addend is considered the main one and the second term is considered as something auxiliary. This is true from the numerical point of view. However, the penalty term is crucial for the examination of theoretical issues such as the existence of an exact penalty representation and/or the calmness of the perturbation function.

In [13, Chapter 11], an augmented Lagrangian with a convex augmenting term was introduced for the primal problem of minimizing an extended real-valued function, and, under mild assumptions, strong duality and exact penalty representation were established [13, Theorems 11.59 and 11.61].

Augmented Lagrangians with a nonconvex augmenting function have been intensively studied as well (see [9, 17, 8, 19, 6, 23, 24, 25] and the references therein). Some of these references (e.g., [9, 17, 6, 19]) use abstract convexity tools in their analysis.

Our aim is (i) to present a unified analysis for the examination of nonconvex augmented Lagrangians for a wider family of augmenting terms, and (ii) to express the main definitions and facts describing the augmented Lagrangian theory in terms of abstract convexity tools.

We extend to our general setting main theoretical facts such as the criterion for exact penalty representation based on local growth properties of the perturbation function (Theorem 6.2), the equivalence between calmness and existence of Lagrange multipliers (Corollary 4.1), and the connection between calmness and existence of exact penalty parameters (Proposition 5.3).

In order to deal with our more general augmenting terms, we translate these main theoretical facts into the language of abstract convexity. Our approach is inspired by the works [13, 17, 8, 23, 25, 10].

*Received by the editors December 14, 2005; accepted for publication (in revised form) November 23, 2006; published electronically May 16, 2007.
http://www.siam.org/journals/siopt/18-2/64762.html

†University of Ballarat, Centre for Informatics and Applied Optimization, Victoria, Australia, and School of Mathematics and Statistics, University of South Australia, Mawson Lakes Campus, Mawson Lakes SA 5095, Australia (Regina.Burachik@unisa.edu.au, http://people.unisa.edu.au/Regina.Burachik).

‡University of Ballarat, Centre for Informatics and Applied Optimization, Victoria, Australia (a.rubinov@ballarat.edu.au).

Recall that a function f is called *abstract convex* with respect to a set of *elementary functions* H if f is represented as the upper envelope of a subset of H. For example, if H is the set of all continuous affine functions, then a function is abstract convex if and only if this function is convex, proper, and lower semicontinuous. In order to apply abstract convexity to the problem at hand we need to define a convenient set H of elementary functions. Since we want to study augmented Lagrangians, we need to consider a set of elementary functions that explicitly depend on a parameter $r > 0$; this parameter is to be identified with a penalty parameter.

Our set of elementary functions is constructed by combining two families of functions. One of these families consists of the "linear" or "ordinary" augmenting terms, and the other provides the "penalty" terms. Our resulting augmenting term includes, as a particular case, the ones studied in [13, 17, 8, 23, 25].

First we prove, in Proposition 4.1, that our augmented Lagrangian scheme has no duality gap. Second, in Proposition 4.2, we establish existence of abstract subgradients of the perturbation function, assuming a local validity of the (abstract) subgradient inequality. Third, we extend the notion of calmness (see Definition 4.1) to our setting and prove in Corollary 4.1 that this new concept is equivalent to the existence of abstract subgradients of the perturbation function at 0. Fourth, we introduce a level boundedness assumption on the problem, which guarantees lower semicontinuity of the perturbation function. We apply these results to the constrained optimization problem and we prove that our concept of calmness implies the existence of an exact penalty parameter. Fifth, we express all previous results in terms of the Lagrangian scheme. More precisely, we prove in Theorem 6.1 the well-definedness of the central path associated with the problem, and we prove optimality of all primal weak accumulation points. The latter result is an extension of [8, Theorem 2.1] and [23, Theorem 3.1] to our general family of augmenting terms. In Theorem 6.2 we establish a criterion for exact penalty representation based on a local behavior of the perturbation function, extending to our setting [8, Theorem 3.1] and [23, Theorem 4.1].

The structure of the paper is as follows: Section 2 contains some preliminaries from abstract convexity and abstract Lagrangians. Generalized augmented Lagrangians are defined in section 3. Abstract subdifferential and approximate abstract subdifferentials of a lower semicontinuous function are discussed in section 4, where we prove zero duality gap and analyze existence of (abstract) subgradients of the perturbation function. Also in this section we establish equivalence between existence of Lagrange multipliers and calmness of the problem. In section 5 we prove that a level-bounded problem guarantees lower semicontinuity of the perturbation function. Also in this section we apply our results to a general constrained optimization problem. In section 6 our previous results (which are expressed in terms of abstract convexity tools) are applied to our augmented Lagrangian scheme. Finally, in section 7 we discuss the application of our duality scheme to a semi-infinite stochastic programming problem.

We use the following notation: $\mathbb{R} = (-\infty, +\infty)$ is the real line; $\bar{\mathbb{R}} = [-\infty, +\infty]$ is the extended real line; $\mathbb{R}_{+\infty} = \mathbb{R} \cup \{+\infty\}$.

Let $f : X \to \bar{\mathbb{R}}$, $g : X \to \bar{\mathbb{R}}$ be functions defined on a set X. Then the inequality $f \leq g$ means that $f(x) \leq g(x)$ for all $x \in X$.

2. Preliminaries.

2.1. Abstract convexity and generalized conjugation. All definitions and statements from this subsection can be found in [18, 21].

Let X be a set and let H be a set of functions $h : X \to \bar{\mathbb{R}}$. Let $f : X \to \bar{\mathbb{R}}$. The set $\operatorname{supp}(f, H) = \{h \in H \mid h \leq f\}$ is called the *support set* of f with respect to H.

The function $\mathrm{co}_H f : X \to \bar{\mathbb{R}}$ defined by $\mathrm{co}_H f(x) = \sup\{h(x) \mid h \in \mathrm{supp}\ (f, H)\}$ is called the H-convex hull of f. A function $f : X \to \bar{\mathbb{R}}$ is called abstract convex with respect to H (H-convex) at a point $x \in X$ if there exists a set $U \subseteq \mathrm{supp}\ (f, H)$ such that $f(x) = \sup\{h(x) \mid h \in U\}$. It is clear that f is H-convex at x if and only if $f(x) = \mathrm{co}_H f(x)$. If f is H-convex at each point $x \in X$, then f is called H-convex on X. Classical convexity is equivalent to abstract convexity with respect to the set A of continuous affine functions. More exactly, if X is a Banach space, then a lower semicontinuous function $f : X \to \mathbb{R}_{+\infty}$ is convex if and only if f is A-convex. Similarly we can define abstract concave with respect to H (H-concave) functions.

Let L be a set of functions defined on a set X. Functions $h_{l,\gamma}$ of the form $h_{l,\gamma}(x) = l(x) - \gamma,\ x \in X$, with $l \in L$ and $\gamma \in \mathbb{R}$ are called L-affine. Denote by H_L the set of all L-affine functions.

Let (X, L) be a pair of sets with a coupling function $\rho : X \times L :\to \mathbb{R}$. The function ρ allows us to consider X as a set of functions $x(\cdot)$ defined on L and L as a set of functions $l(\cdot)$ defined on X. Here

$$x(l) = \rho(x, l) \quad (l \in L), \qquad l(x) = \rho(x, l) \quad (x \in X).$$

Denote by F_X the union of the set of all functions $f : X \to \mathbb{R}_{+\infty}$ and the function $-\infty$, where $-\infty(x) = -\infty$ for all $x \in X$. Let $f \in F_X$. The function

$$f^\rho(l) = \sup_{x \in X}(\rho(x, l) - f(x)), \qquad l \in L,$$

is called the Fenchel–Moreau conjugate of f. This function is H_X-convex. It is easy to check that $f^\rho \in F_L$. The function

$$f^{\rho\rho}(x) = \sup_{l \in L}(\rho(x, l) - f^\rho(l))$$

is called the Fenchel–Moreau biconjugate to f. This function is H_L-convex.

The classical result of abstract convex analysis (the Fenchel–Moreau theorem) states that $f \in F_X$ is abstract convex with respect to H_L at a point x if and only if $f(x) = f^{\rho\rho}(x)$.

Let (X, L) be a pair of sets with a finite coupling function ρ. Let f be H_L convex and $x_0 \in \mathrm{dom}\, f := \{x \in X \mid f(x) < +\infty\}$. The set

$$\begin{aligned}\partial_\rho f(x_0) &= \{l \in L \mid f(x) \geq f(x_0) - l(x_0) + l(x), \quad x \in X\} \\ &= \{l \in L \mid f(x) \geq f(x_0) - \rho(x_0, l) + \rho(x, l), \quad x \in X\}\end{aligned}$$

is called the ρ-subdifferential of f at a point x_0. The ρ-subdifferential $\partial_\rho f(x_0)$ is nonempty if and only if $f(x_0) = \max\{h(x_0) \mid h \in \mathrm{supp}\ (f, H_L)\}$. Elements of $\partial_\rho f(x_0)$ are called ρ-subgradients (abstract subgradients) of f at x_0.

The *approximate* abstract subgradients are inspired by the concept of ε-subgradients introduced in [2]. Fix $\varepsilon \geq 0$. We say that $l \in \partial_{\rho,\varepsilon} f(x_0)$ whenever

$$f(x) \geq f(x_0) - \rho(x_0, l) + \rho(x, l) + \varepsilon \quad \text{for all} \quad x \in X.$$

2.2. Abstract Lagrangians. Definitions and results presented in this subsection are found in [21, 19] (see also the references therein). Let X, Z be reflexive Banach spaces. We consider the optimization problem

$$\text{(P)} \qquad \text{minimize} \ \ \varphi(x) \quad \text{subject to } x \text{ in } X,$$

where the function $\varphi : X \to \mathbb{R}_{+\infty}$ is an extended real-valued function. We will assume that the function φ is proper, that is, $\operatorname{dom}\varphi \neq \emptyset$.

A function $f : X \times Z \to \bar{\mathbb{R}}$ is called a *dualizing parameterization function* for φ if $f(x,0) = \varphi(x)$ for all $x \in X$. This parameterization induces the *perturbation function* given by

$$\beta(z) := \inf_{x\in X} f(x,z).$$

Our assumption on φ forces $\beta(0) < +\infty$.

We introduce duality for problem (P) in the following way. Let Ω be a set of parameters (or dual variables) and consider a coupling function $\rho : Z \times \Omega \to \mathbb{R}$. The *Lagrangian-type function* for problem (P) induced by this coupling function ρ is defined as

$$l(x,\omega) := \inf_{z\in Z}\{f(x,z) - \rho(z,\omega)\}. \tag{2.1}$$

Let $f_x(z) = f(x,z)$. We have

$$l(x,\omega) = -\sup_{z\in Z}\{\rho(z,\omega) - f_x(z)\} = -f_x^{\rho}(\omega).$$

The above expression establishes a relationship between the Lagrangian and the Fenchel–Moreau conjugate to f. This relationship is the key to using conjugations to study duality.

Associated with the Lagrangian function above, we define the *generalized dual function* $q : \Omega \to \bar{\mathbb{R}}$ in a canonical way:

$$q(\omega) := \inf_{x\in X} l(x,\omega) = \inf_{x\in X}(-f_x^{\rho}(\omega)). \tag{2.2}$$

With these definitions, the *generalized dual problem* becomes

$$\sup_{\omega\in\Omega} q(\omega).$$

Assume in what follows that the coupling function ρ verifies $\rho(0,\omega) = 0$ for all $\omega \in \Omega$. Then by (2.1) and (2.2) we get

$$\sup_{\omega} q(\omega) \leq \inf_{x} \varphi(x), \tag{2.3}$$

which is known as the *weak duality property*. When also the opposite inequality holds in (2.3), we say that the *zero duality gap property* holds for the Lagrangian l. The definitions give $\sup_\omega q(\omega) = \beta^{\rho\,\rho}(0)$, where $\beta^{\rho\,\rho}$ is the biconjugate to β with respect to ρ. Altogether, we can write the zero duality gap property as

$$\beta(0) = \beta^{\rho\,\rho}(0).$$

An element $\bar{\omega} \in \Omega$ is called an *exact Lagrange multiplier* if $\beta(0) = q(\bar{\omega}) = \max_{\omega\in\Omega} q(\omega)$. The following result is well known (see, e.g., [19, Theorem 5.2]).

PROPOSITION 2.1. *An element $\bar{\omega} \in \Omega$ is an exact Lagrange multiplier if and only if $\bar{\omega}$ is a ρ-subgradient of β at zero.*

3. Generalized augmented Lagrangians. An abstract Lagrangian is called augmented Lagrangian if $\Omega = Y \times \mathbb{R}_+$, where $Y = Z^*$ is a conjugate to Z space and for $\omega = (y, r) \in \Omega$ we have $\rho(z, (y, r)) = y(z) - r\sigma(z)$. Here σ is an *augmenting* function. It is assumed that σ enjoys some properties; in particular $\sigma(0) = 0$ and $\sigma(z) > 0$ for $z \neq 0$. From the numerical point of view the presence of the linear term $y(z)$, which can be considered as a coupling function generating the classical Lagrangian, is important. However, the role of this term is not essential in the examination of some important theoretical questions, such as the zero duality gap property or the existence of exact penalty parameters. We show that the key role in the study of these questions is played by the penalty term $r\sigma(z)$. In order to underline this key role, we consider a generalized augmented Lagrangian, which contains the augmented Lagrangian described above as a very special case. In order to introduce this generalized augmented Lagrangian we need a function $p : \mathbb{R}^2 \to \mathbb{R}$ which possesses the following properties:

(a) $p(0, 0) = 0$;

(b) there exists a strictly increasing function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ such that $\psi(0) = 0$ and for each $a \in \mathbb{R}$ and $b_1, b_2 \in \mathbb{R}$ with $b_1 \geq b_2$ it holds that

$$p(a, b_1) - p(a, b_2) \geq \psi(b_1 - b_2). \tag{3.1}$$

We will always assume that p verifies conditions (a) and (b).

The simplest example of a function p with such properties is $p(a, b) = a + b$. More generally, we can have $p(a, b) = g(a) + h(b)$ with $g(0) = h(0) = 0$ and h a strictly increasing function. We now give a less trivial example. Let $p(a, b) = g(a)h(b)$, where $g(a) \geq 1$ for all $a \in \mathbb{R}$, and let $h : \mathbb{R} \to \mathbb{R}$ be a differentiable function such that $h'(x) \geq \gamma > 0$ for all x and $h(0) = 0$. Let $b_1 > b_2$. Then

$$p(a, b_1) - p(a, b_2) = g(a)(h(b_1) - h(b_2)) \geq h(b_1) - h(b_2) \geq \gamma(b_1 - b_2),$$

so (3.1) holds with $\psi(b) = \gamma b$.

We need to have two sets U, Y and two families of real-valued functions indexed by these sets: $\{\sigma_u\}_{u \in U}$ and $\{\nu_y\}_{y \in Y}$. We assume that the family $\{\sigma_u\}_{u \in U}$ verifies the following properties:

(U_1) $\sigma_u : Z \to \mathbb{R}_{+\infty}$ are proper, weakly lower semicontinuous for all $u \in U$.

(U_2) $\sigma_u(0) = 0$ and $\sigma_u(z) > 0$ for all $z \neq 0$.

(U_3) For all $u \in U$ there exists $K_u > 0$ such that the level set $\{z \mid \sigma_u(z) \leq K_u\}$ is bounded.

(U_4) For each neighborhood $V \subset Z$ of 0 and for each $u \in U$ we have that

$$\inf_{z \notin V} \sigma_u(z) > 0.$$

Remark 3.1. The function $\mu_u = -\sigma_u$, where σ_u enjoys U_2, U_4 and is continuous, is usually called a peak at zero (see [19]). Penot [10] used the term *potential* for functions similar to σ. In [24], a function that verifies properties U_1, U_2, and U_4 is said to be a *valley at* 0 in X.

We assume that the family $\{\nu_y\}_{y \in Y}$ verifies

(Y_1) $\nu_y(0) = 0$ for all $y \in Y$;

(Y_2) $\nu_y : Z \to \mathbb{R}$ is weakly upper semicontinuous.

Remark 3.2. Each member of the family $\{\sigma_u\}_{u \in U}$ represents a "penalty" term in the augmented Lagrangian, while the family $\{\nu_y\}_{y \in Y}$ consists of the "linear" or "ordinary" augmenting terms.

We now introduce a set of dual variables Ω and a coupling function $\rho : Z \times \Omega \to \mathbb{R}$ which will allow us to combine the elements of both families by using the function p:

$$\Omega := \mathbb{R}_+ \times Y \times U; \tag{3.2}$$

$$\rho(z, \omega) = \rho(z, (r, y, u)) := p(\nu_y(z), -r\sigma_u(z)). \tag{3.3}$$

Example 3.1. Let $\|\cdot\|$ be a norm in the Banach space Z. Take $Y := Z^*$ and $U = (0, +\infty)$. Consider the families $\{\sigma_u\}_{u\in U}$ and $\{\nu_y\}_{y\in Y}$ defined as $\nu_y(z) := y(z)$ and $\sigma_u(z) = \|z\|^u$. It is clear from the definitions that these families verify conditions U_1–U_4. Assume now that $U = (0, 1)$ and consider the family $\{\sigma_u\}_{u\in U}$ given by $\sigma_u(z) = \|z\|^u$ for all z with $\|z\| \le 1$ and $\sigma_u(z) = 1$ otherwise. It has been proved in [24, Lemma 2.1] that the family $\{\sigma_u\}_{u\in U}$ verifies U_1–U_4.

Remark 3.3. Assume that U consists of one element u and call $\sigma_u =: \sigma$. Then Ω consists of vectors $\omega = (r, y)$ with $r \in \mathbb{R}_+$ and $y \in Y$. Let β be a function defined on Z. Then supp $(\beta, H_\Omega) \neq \emptyset$ means that there exists $c \in \mathbb{R}$ and $(r, y) \in \mathbb{R}_+ \times Y$ such that $\beta(z) \ge \nu_y(z) - r\sigma(z) - c$ for all $z \in Z$. This property is related to the concept of growth condition studied in [10]. In [25] this inequality is considered when it holds outside some neighborhood of zero, and it is also called a growth condition.

Remark 3.4. Let $p(a, b) = a + b$ and let $Y = Z^*$ be the dual of the Banach space Z. Assume that U consists of one element u and call $\sigma_u =: \sigma$. Then the augmented Lagrangian $l(x, \omega)$ with $\omega = (r, y)$ has the form

$$l(x, \omega) = \inf_{z\in Z}\{f(x, z) - y(z) + r\sigma(z)\},$$

so the Lagrangian obtained above reduces to the classical augmented Lagrangian defined by the augmenting function σ (see [13, Chapter 11]).

Remark 3.5. Assume that Y consists of one element y and $\nu_y = 0$. Then Ω consists of vectors $\omega = (r, u)$ with $r \in \mathbb{R}_+$ and $u \in U$. In such a case the so-obtained Lagrangian $l(x, \omega) = \inf_{z\in Z}\{f(x, z) + r\sigma_u(z)\}$ is interpreted as a generalized penalty function.

4. Abstract subdifferential and approximate subdifferentials. In this section we examine abstract subdifferential and approximate subdifferential of a lower semicontinuous function β with respect to a coupling function defined by (3.3). Let Ω be a set defined by (3.2). Recall that the set of Ω-affine functions H_Ω consists of functions h of the form $h(z) = \rho(z, (r, y, u)) - c = p(\nu_y(z), -r\sigma_u(z)) - c$, where $\omega = (r, y, u) \in \Omega$ and $c \in \mathbb{R}$.

We consider only functions β such that the support set supp $(\beta, H_\Omega) \neq \emptyset$. This means that there exist $\omega = (r, y, u) \in \Omega$ and $c \in \mathbb{R}$ such that $\beta(z) \ge p(\nu_y(z), -r\sigma_u(z)) - c$ for all $z \in Z$. By definition of β^ρ, the latter inequality holds if and only if $\omega = (r, y, u) \in \operatorname{dom}\beta^\rho$. So our basic assumption supp $(\beta, H_\Omega) \neq \emptyset$ is equivalent to $\operatorname{dom}\beta^\rho \neq \emptyset$.

THEOREM 4.1. *Let X, Z be reflexive Banach spaces. Take $\rho : Z \times \Omega \to \mathbb{R}$ as in* (3.3) *for families of functions verifying U_1–U_3 and Y_1–Y_2. Assume also that $p(\nu_y(\cdot), -r\sigma_u(\cdot))$ is upper semicontinuous for every $(r, y, u) \in \Omega$. Let $\beta : Z \to \mathbb{R}_{+\infty}$ be a lower semicontinuous function such that $\beta(0) < +\infty$ and* supp $(\beta, H_\Omega) \neq \emptyset$. *Then there exists $r_0 > 0$ and $(\bar u, \bar y) \in U \times Y$ such that for all $r > r_0$ the following hold:*

(i) *there exists $z_r \in Z$ such that $(r, \bar y, \bar u) \in \partial_\rho\beta(z_r)$;*

(ii) $(r, \bar y, \bar u) \in \partial_{\rho,\varepsilon_r}\beta(0)$,

where $\varepsilon_r := \beta(0) - \beta(z_r) + p(\nu_{\bar y}(z_r), -r\sigma_{\bar u}(z_r)) \geq 0$. *Furthermore,* $\lim_{r\to\infty} \varepsilon_r = 0$.

Proof. Since supp $(\beta, H_\Omega) \neq \emptyset$, there exist $\bar\omega := (r, \bar y, \bar u) \in \Omega$ and $\bar c \in \mathbb{R}$ such that

$$p(\nu_{\bar y}(z), -\bar r\sigma_{\bar u}(z)) - \bar c \leq \beta(z) \quad \text{for all} \quad z \in Z.$$

For any $r \geq 0$ call $m(r) := \inf_{z\in Z}\{\beta(z) - p(\nu_{\bar y}(z), -r\sigma_{\bar u}(z))\}$. The equalities $\nu_{\bar y}(0) = \sigma_{\bar u}(0) = 0$ and $p(0,0) = 0$ imply $m(r) \leq \beta(0)$. Fix $r > \bar r > 0$. Since p is increasing on the second argument and $\sigma_u \geq 0$ it follows that

$$\begin{aligned} -\infty < m(\bar r) &= \inf_{z\in Z}\{\beta(z) - p(\nu_{\bar y}(z), -\bar r\sigma_{\bar u}(z))\} \\ &\leq \inf_{z\in Z}\{\beta(z) - p(\nu_{\bar y}(z), -r\sigma_{\bar u}(z))\} = m(r) \leq \beta(0). \end{aligned}$$

Take a sequence $\{z_n\}$ such that $\lim_{n\to\infty} \beta(z_n) - p(\nu_{\bar y}(z_n), -r\sigma_{\bar u}(z_n)) = m(r)$. Call $\theta_n := \beta(z_n) - p(\nu_{\bar y}(z_n), -r\sigma_{\bar u}(z_n))$. There exists n_0 such that $\theta_n < \beta(0) + 1$ for all $n \geq n_0$. We can write for $n \geq n_0$

$$\begin{aligned} m(\bar r) &\leq \beta(z_n) - p(\nu_{\bar y}(z_n), -\bar r\sigma_{\bar u}(z_n)) \\ &= \beta(z_n) - p(\nu_{\bar y}(z_n), -r\sigma_{\bar u}(z_n)) + p(\nu_{\bar y}(z_n), -r\sigma_{\bar u}(z_n)) - p(\nu_{\bar y}(z_n), -\bar r\sigma_{\bar u}(z_n)) \\ &= \theta_n + (p(\nu_{\bar y}(z_n), -r\sigma_{\bar u}(z_n)) - p(\nu_{\bar y}(z_n), -\bar r\sigma_{\bar u}(z_n))). \end{aligned}$$

Let $d_1 = -\bar r\sigma_{\bar u}(z_n), d_2 = -r\sigma_{\bar u}(z_n)$. Then $d_1 \geq d_2$. Due to (3.1), we have

$$p(\nu_{\bar y}(z_n), d_1) - p(\nu_{\bar y}(z_n), d_2) \geq \psi(d_1 - d_2) = \psi((r - \overline{r})\sigma_{\bar u}(z_n)).$$

Hence for $n \geq n_0$ we can write

$$m(\overline{r}) \leq \theta_n - \psi((r - \overline{r})\sigma_{\bar u}(z_n)) \leq \beta(0) + 1 - \psi((r - \overline{r})\sigma_{\bar u}(z_n)).$$

Rearranging the above expression we get for $n \geq n_0$

$$\psi((r - \bar r)\sigma_{\bar u}(z_n)) \leq \beta(0) + 1 - m(\overline{r}),$$

so it holds that

$$\sigma_{\bar u}(z_n) \leq \frac{\psi^{-1}(\beta(0) + 1 - m(\bar r))}{r - \bar r}$$

for $n \geq n_0$. Let $K_{\bar u}$ be as in condition U_3. Take

$$r_0 := \bar r + \frac{\psi^{-1}(\beta(0) + 1 - m(\bar r))}{K_{\bar u}}.$$

Then for all $r > r_0$ and all $n \geq n_0$ we get $\sigma_{\bar u}(z_n) \leq K_{\bar u}$. By condition U_3, this implies that $\{z_n\}_{n\geq n_0}$ is bounded. Therefore there exists a subsequence $\{z_{n_j}\}$ of $\{z_n\}$ weakly converging to some z_r. By lower semicontinuity of all functions involved, we can write

$$\begin{aligned} \beta(z_r) - p(\nu_{\bar y}(z_r), -r\sigma_{\bar u}(z_r)) &\leq \liminf_j \beta(z_{n_j}) - p(\nu_{\bar y}(z_{n_j}), -r\sigma_{\bar u}(z_{n_j})) \\ &= \liminf_j \theta_{n_j} = m(r). \end{aligned}$$

By definition of $m(r)$, the above inequality implies that for all $r > r_0$ and for all z we have

$$\beta(z_r) - p(\nu_{\bar y}(z_r), -r\sigma_{\bar u}(z_r)) \leq \beta(z) - p(\nu_{\bar y}(z), -r\sigma_{\bar u}(z))$$

or, equivalently,

$$\beta(z) \geq \beta(z_r) + \rho(z,(r,\bar{y},\bar{u})) - \rho(z_r,(r,\bar{y},\bar{u})). \tag{4.1}$$

This means that $\bar{\omega} = (r,\bar{y},\bar{u}) \in \partial_\rho \beta(z_r)$. This completes the proof of (i).

Let us prove (ii). The fact that $\varepsilon_r \geq 0$ follows readily from (4.1) for $z = 0$. Since $\rho(0,\omega) = 0$, the inclusion in condition (ii) is equivalent to checking that

$$\beta(z) - \beta(0) - p(\nu_{\bar{y}}(z), -r\sigma_{\bar{u}}(z)) + \varepsilon_r \geq 0 \quad \text{for all } \ z \in Z. \tag{4.2}$$

Indeed, since $\bar{\omega} \in \partial_\rho \beta(z_r)$ we have

$$\begin{aligned} 0 &\leq \beta(z) - p(\nu_{\bar{y}}(z), -r\sigma_{\bar{u}}(z)) - \beta(z_r) + p(\nu_{\bar{y}}(z_r), -r\sigma_{\bar{u}}(z_r)) \\ &= \beta(z) - \beta(0) - p(\nu_{\bar{y}}(z), -r\sigma_{\bar{u}}(z)) - \beta(z_r) + \beta(0) + p(\nu_{\bar{y}}(z_r), -r\sigma_{\bar{u}}(z_r)) \\ &= \beta(z) - \beta(0) - p(\nu_{\bar{y}}(z), -r\sigma_{\bar{u}}(z)) + \varepsilon_r, \end{aligned}$$

which establishes (4.2). For proving that $\lim_{r\to\infty} \varepsilon_r = 0$ we need to establish the following fact.

Fact 1. The function $r \mapsto z_r$ $(r > r_0)$ is bounded and converges weakly to 0 as $r \to +\infty$.

Indeed, assume Fact 1 is true. Without loss of generality we can assume that $r_0 \geq 1$. Since p is increasing on the second argument and $\nu_{\bar{y}}(0) = \sigma_{\bar{u}}(0) = 0$ it follows from (4.1) with $z = 0$ that

$$\begin{aligned} \beta(0) &\geq \limsup_{r\to+\infty}(\beta(z_r) - p(\nu_{\bar{y}}(z_r), -r\sigma_{\bar{u}}(z_r))) \\ &\geq \liminf_{r\to+\infty}(\beta(z_r) - p(\nu_{\bar{y}}(z_r), -r\sigma_{\bar{u}}(z_r))) \\ &\geq \liminf_{r\to+\infty} \beta(z_r) + \liminf_{r\to+\infty} -p(\nu_{\bar{y}}(z_r), -r\sigma_{\bar{u}}(z_r)) \geq \beta(0), \end{aligned}$$

where we used the upper semicontinuity assumption on p and lower semicontinuity of β. Therefore $\beta(0) = \lim_{r\to+\infty}(\beta(z_r) - p(\nu_{\bar{y}}(z_r), -r\sigma_{\bar{u}}(z_r))$. Now the definition of ε_r shows that $\lim_{r\to\infty} \varepsilon_r = 0$. So let us prove Fact 1.

Proof of Fact 1. From the definition of $m(\overline{r})$ we can write for every $r > r_0$

$$\begin{aligned} m(\overline{r}) &\leq \beta(z_r) - p(\nu_{\bar{y}}(z_r), -\overline{r}\sigma_{\bar{u}}(z_r)) \\ &= \beta(z_r) - p(\nu_{\bar{y}}(z_r), -r\sigma_{\bar{u}}(z_r)) + p(\nu_{\bar{y}}(z_r), -r\sigma_{\bar{u}}(z_r)) - p(\nu_{\bar{y}}(z_r), -\overline{r}\sigma_{\bar{u}}(z_r)). \end{aligned}$$

Since $r > \overline{r}$ it follows that

$$p(\nu_{\bar{y}}(z_r), -\overline{r}\sigma_{\bar{u}}(z_r)) - p(\nu_{\bar{y}}(z_r), -r\sigma_{\bar{u}}(z_r)) \geq \psi((r-\overline{r})\sigma_{\bar{u}}(z_r)),$$

so

$$m(\overline{r}) \leq \beta(z_r) - p(\nu_{\bar{y}}(z_r), -r\sigma_{\bar{u}}(z_r)) - \psi((r-\overline{r})\sigma_{\bar{u}}(z_r)).$$

We also have

$$\beta(z_r) - p(\nu_{\bar{y}}(z_r), -r\sigma_{\bar{u}}(z_r)) = m(r) \leq \beta(0).$$

Therefore

$$\psi((r-\overline{r})\sigma_{\bar{u}}(z_r)) \leq \beta(z_r) - p(\nu_{\bar{y}}(z_r), -r\sigma_{\bar{u}}(z_r)) - m(\overline{r}) \leq \beta(0) - m(\overline{r}).$$

It follows from this inequality that

$$\sigma_{\bar{u}}(z_r) \le \frac{\psi^{-1}(\beta(0) - m(\overline{r}))}{r - \overline{r}}. \tag{4.3}$$

Using the definition of r_0 and the fact that $r > r_0$ we conclude that

$$\sigma_{\bar{u}}(z_r) \le K_{\bar{u}} \frac{\psi^{-1}(\beta(0) - m(\overline{r}))}{\psi^{-1}(\beta(0) + 1 - m(\overline{r}))} \le K_{\bar{u}}.$$

So the function $r \mapsto z_r$ $(r \ge r_0)$ is bounded. Now let $\overline{z}$ be a weak accumulation point of this function as $r \to +\infty$. Expression (4.3) yields

$$\begin{aligned} 0 \le \sigma_{\bar{u}}(\overline{z}) &\le \liminf_{r\to\infty} \sigma_{\bar{u}}(z_r) \\ &\le \liminf_{r\to\infty} \frac{\psi^{-1}(\beta(0) - m(\overline{r}))}{r - \overline{r}} = 0. \end{aligned}$$

Hence $\overline{z} = 0$. Since this is the unique possible accumulation point, the function converges weakly to 0 as $r \to +\infty$. The proof of (ii) is now complete. □

Now we use the previous result to prove that the function β is abstract convex with respect to H_Ω at zero. In other words, our duality scheme has zero duality gap.

PROPOSITION 4.1. *The equality*

$$\beta(0) = \beta^{\rho\,\rho}(0)$$

holds under the assumptions of Theorem 4.1.

Proof. Since $\rho(0, \omega) = 0$ for all $\omega \in \Omega$, it is enough to prove that $\beta(0) \le \beta^{\rho\,\rho}(0)$. With the notation of Theorem 4.1(i), there exists $r_0 > 0$ such that for all $r \ge r_0$ we have $(r, \bar{y}, \bar{u}) \in \partial_\rho \beta(z_r)$. It is easy to see from the definition of abstract subdifferential that this inclusion implies $\beta^\rho(r, \bar{y}, \bar{u}) = p(\nu_{\bar{y}}(z_r), -r\sigma_{\bar{u}}(z_r)) - \beta(z_r)$. By Fact 1 in the proof of Theorem 4.1, $\{z_r\}$ weakly converges to zero as $r \to +\infty$. Hence,

$$\beta^{\rho\,\rho}(0) = \sup_{\omega\in\Omega} \rho(\omega, 0) - \beta^\rho(\omega) \ge -\beta^\rho(r, \bar{y}, \bar{u}) = -p(\nu_{\bar{y}}(z_r), -r\sigma_{\bar{u}}(z_r)) + \beta(z_r).$$

Altogether we have

$$\beta^{\rho\,\rho}(0) \ge \liminf_{r\to\infty} \beta(z_r) - p(\nu_{\bar{y}}(z_r), -r\sigma_{\bar{u}}(z_r)) \ge \beta(0).$$

The proof is complete. □

The definition of abstract subdifferential shows that an abstract subgradient has a global nature in the sense that we need to check a certain inequality for all points $z \in Z$ in order to check that an element $\omega \in \Omega$ is a subgradient of a function β defined on Z at a point $z_0 \in Z$. It is interesting to establish conditions under which we can check the required inequality only in a certain neighborhood of a point z. The next statement shows that this local property is valid for the coupling function ρ defined by (3.3) for large enough r.

PROPOSITION 4.2. *Assume all hypotheses of Theorem* 4.1 *hold (in particular,* $\operatorname{dom}\beta^\rho \ne \emptyset$*). Take* $(r, \bar{y}, \bar{u}) \in \operatorname{dom}\beta^\rho$ *and assume that condition* U_4 *is satisfied by* $\sigma_{\overline{u}}$*. If there exists a neighborhood* V *of* 0 *and* $r' > 0$ *such that for all* $z \in V$ *it holds that*

$$\beta(z) \ge \beta(0) + p(\nu_{\bar{y}}(z), -r'\sigma_{\bar{u}}(z)), \tag{4.4}$$

then there exists $r^* \geq r'$ *such that* $(r, \bar{y}, \bar{u}) \in \partial_\rho \beta(0)$ *for all* $r > r^*$.

Proof. Take r_0 as in Theorem 4.1. By Theorem 4.1(ii) we have that $(r, \bar{y}, \bar{u}) \in \partial_{\rho,\varepsilon_r}\beta(0)$ for all $r > r_0$, with $\lim_{r\to\infty} \varepsilon_r = 0$. Fix $\varepsilon_0 > 0$. Since ε_r tends to 0, there exists $r_1 > r_0$ such that $\varepsilon_0 \geq \varepsilon_r$ for all $r > r_1$. By condition U_4, there exists $c(V, \bar{u}) > 0$ such that

$$c(V, \bar{u}) < \inf_{z \notin V} \sigma_{\bar{u}}(z).$$

Fix $r_2 > r_1$ and choose

$$r^* > \max\left\{r', \frac{\psi^{-1}(\varepsilon_0)}{c(V, \bar{u})} + r_2\right\}. \tag{4.5}$$

Since $r^* > r'$ we have by assumption (4.4) and condition (b) on p that the inequality

$$\beta(z) \geq \beta(0) + p(\nu_{\bar{y}}(z), -r\sigma_{\bar{u}}(z))$$

holds for all $z \in V$ and all $r > r^*$. Assume now that $z \notin V$ and $r > r^*$. Because $r_2 > r_0$ we have that $(r_2, \bar{y}, \bar{u}) \in \partial_{\rho,\varepsilon_{r_2}}\beta(0)$. Hence

$$\begin{aligned}\beta(0) - \varepsilon_{r_2} &\leq \beta(z) - p(\nu_{\bar{y}}(z), -r_2\sigma_{\bar{u}}(z)) \\ &= \beta(z) - p(\nu_{\bar{y}}(z), -r^*\sigma_{\bar{u}}(z)) + p(\nu_{\bar{y}}(z), -r^*\sigma_{\bar{u}}(z)) - p(\nu_{\bar{y}}(z), -r_2\sigma_{\bar{u}}(z)).\end{aligned}$$

Since $r^* \geq r_2$, it follows that

$$p(\nu_{\bar{y}}(z), -r_2\sigma_{\bar{u}}(z)) - p(\nu_{\bar{y}}(z), -r^*\sigma_{\bar{u}}(z)) \geq \psi((r^* - r_2)\sigma_{\bar{u}}(z)),$$

so

$$\beta(0) - \varepsilon_{r_2} \leq \beta(z) - p(\nu_{\bar{y}}(z), -r^*\sigma_{\bar{u}}(z)) - \psi((r^* - r_2)\sigma_{\bar{u}}(z)).$$

Since $\sigma_{\bar{u}}(z) > c(V, \bar{u})$ and ψ is increasing we have

$$\beta(0) - \varepsilon_{r_2} \leq \beta(z) - p(\nu_{\bar{y}}(z), -r^*\sigma_{\bar{u}}(z)) - \psi((r^* - r_2)c(V, \bar{u})). \tag{4.6}$$

According to (4.5) we have $\varepsilon_0 < \psi((r^* - r_2)c(V, \bar{u}))$; therefore $-\psi((r^* - r_2)c(V, \bar{u})) \leq -\varepsilon_0 \leq -\varepsilon_{r_2}$, which combined with (4.6) gives

$$\beta(0) - \varepsilon_{r_2} \leq \beta(z) - p(\nu_{\bar{y}}(z), -r^*\sigma_{\bar{u}}(z)) - \varepsilon_{r_2}.$$

Hence we get

$$\beta(0) \leq \beta(z) - p(\nu_{\bar{y}}(z), -r^*\sigma_{\bar{u}}(z))$$

for all $z \notin V$. By condition (b) in the definition of p, the inequality holds for all $r > r^*$. The proof is complete. □

We quoted in Proposition 2.1 the equivalence between the existence of an exact penalty parameter and nonemptiness of the subgradient of the perturbation function β at zero. In the classical framework another equivalent property holds. This property is known as *calmness at zero* of the perturbation function. Under some constraint qualifications (see, e.g., [13, Proposition 8.32]), calmness at zero of β is equivalent to nonemptiness of the subdifferential of β at 0. In our context, we need to introduce a

concept of calmness which is related to the family $\{\sigma_u\}$. This concept reduces to the classical one when $\sigma_u = \|\cdot\|$ for all $u \in U$.

DEFINITION 4.1. *Fix $u \in U$ and let $h : Z \to \bar{\mathbb{R}}$. We say that h is* σ_u-calm at 0 *when*

$$\liminf_{z\to 0,\, z\neq 0} \frac{h(z) - h(0)}{\sigma_u(z)} > -\infty.$$

In order to establish the relation between calmness at zero and abstract subdifferentiability, we need to assume a growth condition on p. Our definition was inspired from [10, Definition 2.6].

DEFINITION 4.2. *Fix $u \in U$ and $y \in Y$. We say that p verifies a* growth condition with respect to the functions ν_y *and* σ_u *when there exists $r = r(y,u) > 0$ such that*

$$\text{(i)} \quad \liminf_{z\to 0,\, z\neq 0} \frac{p(\nu_y(z), -r\sigma_u(z))}{\sigma_u(z)} > -\infty.$$

We say that σ_u is a penalty function *for p and ν_y when*

$$\text{(ii)} \quad \limsup_{z\to 0,\, z\neq 0,\, r\to+\infty} \frac{p(\nu_y(z), -r\sigma_u(z))}{\sigma_u(z)} = -\infty.$$

Example 4.1. Let $p(a,b) = a + b$ and let the families $\{\nu_y\}_y$ and $\{\sigma_u\}_u$ be as in Example 3.1 with $U = (0,1]$. Then both (i) and (ii) of Definition 4.2 hold for every pair σ_u, ν_y.

The proposition below relates Definition 4.2 with nonemptiness of $\partial_\rho\beta(0)$.

PROPOSITION 4.3.

(a) *Assume $\partial_\rho\beta(0)$ is nonempty and take $(\bar r, \bar y, \bar u) \in \partial_\rho\beta(0)$. If p verifies a growth condition with respect to the functions $\nu_{\bar y}$ and $\sigma_{\bar u}$, then β is $\sigma_{\bar u}$-calm at 0.*

(b) *Assume all hypotheses of Theorem* 4.1 *hold, and fix $(\bar r, \bar y, \bar u) \in \operatorname{dom}\beta^\rho$. If β is $\sigma_{\bar u}$-calm at 0 and $\sigma_{\bar u}$ is a penalty function for p and $\nu_{\bar y}$, then the subdifferential $\partial_\rho\beta(0)$ is nonempty.*

Proof. (a) We have that $\partial_\rho\beta(0) \neq \emptyset$ and $(r, \overline{y}, \overline{u}) \in \partial_\rho\beta(0)$. Then $\beta(z) \geq \beta(0) + p(\nu_{\overline{y}}(z), -r\sigma_{\overline{u}}(z))$. This gives

$$\liminf_{z\to 0} \frac{\beta(z) - \beta(0)}{\sigma_{\bar u}(z)} \geq \liminf_{z\to 0} \frac{p(\nu_{\overline{y}}(z), -r\sigma_{\bar u}(z))}{\sigma_{\bar u}(z)} > -\infty,$$

which implies the $\sigma_{\bar u}$-calmness at 0.

(b) Assume now that there exists $\bar u \in U$ such that β is $\sigma_{\bar u}$-calm at 0. Hence there exists $L > 0$ and a neighborhood W of 0 such that for all $z \in W \setminus \{0\}$ we have

$$\frac{\beta(z) - \beta(0)}{\sigma_{\bar u}(z)} > -L.$$

On the other hand, by condition (ii) in Definition 4.2 we have that for the given $\bar u$ there exists $\bar y \in Y$ such that

$$\limsup_{z\to 0,\, z\neq 0,\, r\to+\infty} \frac{p(\nu_{\overline{y}}(z), -r\sigma_{\bar u}(z))}{\sigma_{\bar u}(z)} < -L.$$

In other words, there exists $r_0 > 0$ and a neighborhood W_0 of 0 such that for all $r > r_0$ and all $z \in W_0 \setminus \{0\}$ we have

$$\frac{p(\nu_{\overline{y}}(z), -r\sigma_{\bar u}(z))}{\sigma_{\bar u}(z)} < -L.$$

Altogether, for all $z \in W \cap W_0 \setminus \{0\}$ and all $r > r_0$ we can write

$$\beta(z) - \beta(0) > -L\sigma_{\bar{u}}(z) > p(\nu_{\overline{y}}(z), -r\sigma_{\bar{u}}(z)).$$

So inequality (4.4) holds for a neighborhood of 0. By Proposition 4.2 we conclude that the inequality holds globally, and hence $(r, \overline{y}, \overline{u}) \in \partial_\rho \beta(0)$ for all $r > r_0$. In particular, $\partial_\rho \beta(0) \neq \emptyset$. □

The result below combines both items of the last proposition.

COROLLARY 4.1. *Assume the hypotheses of Theorem* 4.1 *hold with* $(\bar{r}, \bar{y}, \bar{u}) \in \operatorname{dom} \beta^\rho$. *Suppose also that* p, $\nu_{\overline{y}}$, *and* $\sigma_{\overline{u}}$ *verify conditions* (i)–(ii) *of Definition* 4.2. *Then the subdifferential* $\partial_\rho \beta(0)$ *is nonempty if and only if* β *is* $\sigma_{\bar{u}}$*-calm at* 0.

5. Lower semicontinuity of abstract concave functions. Let H be a set of functions defined on a topological space Z and let $\beta : Z \to \mathbb{R}_{-\infty}$ be an H-concave function; that is, there is a set $U \subset H$ such that $\beta(z) = \inf_{h \in U} h(z)$. If H consists of continuous functions, then β is upper semicontinuous. On the other hand, a basic assumption in Theorem 4.1 is lower semicontinuity of β. Therefore, we are interested in assumptions which guarantee the latter property, assuming that H consists of lower semicontinuous functions.

Let $f : X \times Z \to \mathbb{R}_{+\infty}$ be a function and

$$\beta(z) = \inf_{x \in X} f(x, z). \tag{5.1}$$

We can consider β as an abstract concave function with respect to a set H of elementary functions which contains all functions $(f_x)_{x \in X}$, where $f_x(z) = f(x, z)$. We will apply Theorem 4.1 for the function β defined by (5.1). For this we need to describe functions f such that the function β is weakly lower semicontinuous. The following definition will be used.

DEFINITION 5.1. *A function* $f : X \times Z \to \bar{\mathbb{R}}$ *is said to be* weakly level-compact *if for every* $\overline{z} \in Z$ *there exists a weak neighborhood* $W \subset Z$ *of* $\overline{z}$ *such that, for every* $\alpha \in \mathbb{R}$, *the set*

$$\{x \in X \mid f(x, z) \leq \alpha\} \subset \bar{B} \text{ for all } z \in W,$$

where $\bar{B} \subset X$ *is weakly compact.*

Remark 5.1. If in Definition 5.1 we require $\bar{B}$ to be a bounded set, then we recover the definition of uniform level boundedness given in [13, Definition 1.16].

Remark 5.2. Since X is a reflexive Banach space, Alaoglu's theorem implies that a set is weakly compact if and only if it is bounded and weakly closed. Moreover, every bounded sequence has a weakly convergent subsequence.

PROPOSITION 5.1. *Let* $f : X \times Z \to \bar{\mathbb{R}}$ *be weakly lower semicontinuous and weakly level-compact. Then the function* β *defined by* (5.1) *is weakly lower semicontinuous.*

Proof. Assume that β is not weakly lower semicontinuous, which means that there exist a point $\overline{z}$, a net $\{z_i\}_{i \in I}$, and $\varepsilon > 0$ such that

(i) $\overline{z} = (w) - \lim_i z_i$;

(ii) $\liminf_i \beta(z_i) < \beta(\overline{z}) - \varepsilon$.

By Definition 5.1, there exists a weak neighborhood W of $\overline{z}$ such that the set

$$\{x \in X \,:\, f(x, z) \leq \beta(\overline{z}) - \varepsilon\} \subset B \text{ for all } z \in W,$$

where B is weakly compact. Item (i) means that there exists a terminal subset J of I ($J \subset I$ is *terminal* when there exists $i_0 \in I$ such that whenever $j \geq i_0$ we must have

$j \in J$) such that $z_j \in W$ for all $j \in J$. Using this fact in (ii) we get

$$\beta(\overline{z}) - \varepsilon > \sup_{I' \subset I,\, I' \text{ terminal}} \inf_{j \in I'} \beta(z_j) \geq \inf_{j \in J} \beta(z_j).$$

Now take $k_0 \in \mathbb{N}$ such that for all $k \geq k_0$ we have

$$\beta(\overline{z}) - \varepsilon - 1/k > \inf_{j \in J} \beta(z_j) = \inf_{j \in J} \inf_{x \in X} f(x, z_j).$$

For every fixed $k \geq k_0$, we can choose $x_k \in X$, $j_k \in J$ such that $f(x_k, z_{j_k}) < \beta(\overline{z}) - \varepsilon - 1/k$. Because B is weakly compact and the sequence $\{x_k\} \subset B$, there exists a subsequence $\{x_{k_l}\}_{l \in \mathbb{N}}$ weakly converging to some $\overline{x}$. Altogether, the subsequence $\{(x_{k_l}, z_{j_{k_l}})\}_l$ weakly converges to $(\overline{x}, \overline{z})$ and by weak lower semicontinuity of f we get

$$\beta(\overline{z}) \leq f(\overline{x}, \overline{z}) \leq \liminf_l f(x_{k_l}, z_{j_{k_l}}) \leq \liminf_l \beta(\overline{z}) - \varepsilon - 1/k_l = \beta(\overline{z}) - \varepsilon,$$

which is impossible. We conclude that β must be weakly lower semicontinuous. ∎

5.1. Application to constrained minimization and exact penalization. We propose now a duality scheme for a general constrained problem. Given X, Z reflexive Banach spaces, consider the constrained problem

$$\text{(P)} \quad \min_{x \in C} h(x), \tag{5.2}$$

where $C \subset X$ and $\varphi : X \to \mathbb{R}_{+\infty}$. Denote by δ_A the indicator function of the set A, i.e., $\delta_A(x) = 0$ if $x \in A$ and $\delta_A(x) = +\infty$ if $x \notin A$. Problem (5.2) can be reformulated as our original problem (P) with $\varphi(x) := h(x) + \delta_C(x)$. In order to define the perturbed problems, we consider a point to set mapping $D : Z \rightrightarrows X$ such that $D(0) = C$. This induces a duality parameterization $f(x, z) := h(x) + \delta_{D(z)}(x)$, so the perturbed problems become

$$(P_z) \qquad \min_{x \in X} f(x, z) = \min_{x \in D(z)} h(x).$$

Remark 5.3. Assume the original problem is a classical nonlinear programming problem with constraint set $C = \{x \in X \,:\, g_i(x) \leq 0,\, i = 1, \ldots, m\}$, where $g_i : X \to \mathbb{R}_{+\infty}$. We can define perturbed problems using the point-to-set mapping $D : \mathbb{R}^m \rightrightarrows X$ given by

$$D(v) := \{x \in X \,:\, g_i(x) \leq v_i,\, i = 1, \ldots, m\}.$$

This choice of D corresponds to the *canonical perturbations* given in [11, Example 1, eq. (2.8)] or [12, section 6].

We will make the following basic assumptions on problem (P) and its perturbations (P_z).

(H0) D is weakly outer semicontinuous (i.e., the graph of D is weakly closed).

(H1) $h : X \to \mathbb{R}_{+\infty}$ is weakly lower semicontinuous.

Remark 5.4. Note that the mapping D defined in Remark 5.3 verifies (H0) when the constraint functions g_i are weakly lower semicontinuous.

Our first step is to guarantee weak lower semicontinuity of β.

PROPOSITION 5.2. *Let* (H0)–(H1) *hold. Assume that h is coercive, i.e.,* $\lim_{\|x\| \to \infty} h(x) = +\infty$. *Consider* $f(x, z) := h(x) + \delta_{D(z)}(x)$. *Then* $\beta(z) := \inf_{x \in X} f(x, z)$ *is weakly lower semicontinuous.*

Proof. Our first step is to prove that f is weakly level-compact. Indeed, for every $\bar{z} \in Z$ and every $W \subset Z$ such that $\bar{z} \in W$ we can write

$$\begin{aligned} \cup_{z\in W} A_\alpha(z) &= \cup_{z\in W}\{x \in X \,:\, f(x,z) \le \alpha\} \\ &= \cup_{z\in W}\{x \in X \,:\, h(x) \le \alpha \text{ and } x \in D(z)\} \\ &\subset h^{-1}((-\infty,\alpha]). \end{aligned}$$

By (H1) and the coercivity assumption, the set $h^{-1}((-\infty,\alpha])$ is bounded and weakly closed, and therefore weakly compact. Hence, f is weakly level-compact. Let us prove now that f is weakly lower semicontinuous at every $(\overline{x},\overline{z}) \in X \times Z$. Let $\{(x_i,z_i)\}_{i\in I}$ be a net converging weakly to $(\overline{x},\overline{z})$. Denote by $J \subset_T I$ the fact that $J \subset I$ and J is terminal. We can write

$$\begin{aligned} \liminf_{i\in I} f(x_i,z_i) &= \sup_{J\subset_T I} \inf_{j\in J} f(x_j,z_j) \\ &= \sup_{J\subset_T I} \inf_{j\in J,\, x_j\in D(z_j)} h(x_j) \\ &\ge \sup_{J\subset_T I} \inf_{j\in J} h(x_j) \ge h(\bar{x}), \end{aligned}$$

where we used assumption (H1) in the last inequality. On the other hand, by (H0) we have that $\bar{x} \in D(\bar{z})$ and therefore $f(\bar{x},\bar{z}) = h(\bar{x})$. The latter fact and the above expression yield the weak lower semicontinuity of f. The conclusion of the proposition now follows from Proposition 5.1. ∎

The concept of calmness is related with exact penalization (see, e.g., [3, Proposition 6.4.3]). Indeed, we establish next a connection between σ_u-calmness and the existence of an exact penalty parameter. More precisely, we prove that under σ_u-calmness of the perturbation function, every solution of problem (5.2) is also a local solution of the penalized problem, as long as the penalty parameter r is large enough. In other words, σ_u is an exact penalty function for problem (5.2). We need extra assumptions on our Lagrangian function.

(H2) The function ψ in (3.1) verifies

$$\liminf_{t\to 0^+} \frac{\psi(t)}{t} \ge \delta > 0$$

for some $\delta > 0$.

(H3) The function ψ in (3.1) verifies

$$\liminf_{r\to+\infty} \psi(r) = +\infty.$$

Remark 5.5. For instance, $\psi(t) = t^\alpha$ with $0 < \alpha \le 1$ verifies (H2) and (H3). In particular, the choice $p(a,b) = a + b$ corresponds to $\psi(t) = t$ for every $t \ge 0$.

PROPOSITION 5.3. *Assume that* (H0)–(H2) *hold. Let* σ_u *satisfy* U_1–U_3 *and let* Y *be such that* $0 \in Y$ *and* $\nu_0(z) = 0$ *for every* $z \in Z$. *Suppose also that either of the following two assumptions holds:*

(i) ψ *in* (3.1) *verifies* (H3).

(ii) h *is continuous on every solution of problem* (5.2).

If β *is* σ_u*-calm at* 0, *then for every solution* x^* *of problem* (5.2), *there exists* $M > 0$ *such that* x^* *is a local (with respect to the strong topology in* X*) solution of*

$$\min_{x\in X} \tilde{l}(x,(M,u)), \tag{5.3}$$

where $\tilde{l}(x,(M,u)) := l(x,(M,0,u))$.

Proof. As in Proposition 5.2, we take $f(x,z) = h(x) + \delta_{D(z)}(x)$. Note that $\rho(z,(r,0,u)) = p(\nu_0(z), -r\sigma_u(z)) = p(0,-r\sigma_u(z))$, so the augmented Lagrangian corresponding to ρ in the element $(x,(r,0,u))$ has the form

$$\begin{aligned} l(x,(r,0,u)) &= \inf_{z\in Z}\{f(x,z) - \rho(z,(r,0,u))\} \\ &= h(x) + \inf_{z\in D^{-1}(x)}\left[-p(0,-r\sigma_u(z))\right]. \end{aligned} \tag{5.4}$$

Note that $0 \in D^{-1}(x^*)$ because $x^* \in C = D(0)$. Using this fact in the above expression for $x = x^*$ yields

$$l(x^*,(r,0,u)) \le h(x^*). \tag{5.5}$$

By (3.1) for $a = b_1 = 0$ and $b_2 = -r\sigma_u(z) \le 0$ we have that

$$-p(0,-r\sigma_u(z)) \ge \psi(r\sigma_u(z)) \ge 0. \tag{5.6}$$

Combining (5.6) with (5.4) we get

$$\begin{aligned} l(x,(r,0,u)) &= h(x) + \inf_{z\in D^{-1}(x)}\left[-p(0,-r\sigma_u(z))\right] \\ &\ge h(x) + \inf_{z\in D^{-1}(x)}\psi(r\sigma_u(z)) \ge h(x). \end{aligned} \tag{5.7}$$

From (5.7) and (5.5) we obtain $l(x^*,(r,0,u)) = h(x^*)$. Assume that the conclusion of the proposition is not true. This implies that for every $k \in \mathbb{N}$ there exists $x_k \in X$ such that $\|x^k - x^*\| < 1/k$ and

$$\tilde{l}(x_k,(k,u)) = h(x_k) + \inf_{z\in D^{-1}(x_k)}\left[-p(0,-k\sigma_u(z))\right] < \tilde{l}(x^*,(k,u)) = h(x^*). \tag{5.8}$$

Call $a_k := \inf_{z\in D^{-1}(x_k)}\left[-p(0,-k\sigma_u(z))\right]$. We claim that $a_k > 0$ for all k. Indeed, if there exists k_0 such that $a_{k_0} = 0$, then we can find a sequence $\{z_j\} \subset D^{-1}(x_{k_0})$ such that $\lim_{j\to\infty} -p(0,-k_0\sigma_u(z_j)) = 0$. But the latter can hold only when $\lim_{j\to\infty}\sigma_u(z_j) = 0$. Indeed, suppose that for some $b > 0$ there is a subsequence $\{z_{j_l}\}_l$ such that $\sigma_u(z_{j_l}) > b$ for all $l \in \mathbb{N}$. Using (5.6) and the fact that ψ is strictly increasing we get

$$0 = \lim_{l\to\infty} -p(0,-k_0\sigma_u(z_{j_l})) \ge \lim_{l\to\infty}\psi(k_0\sigma_u(z_{j_l})) \ge \psi(b) > 0,$$

where we also used $\psi(0) = 0$ in the last inequality. The above expression entails a contradiction, and hence we must have $\lim_{j\to\infty}\sigma_u(z_j) = 0$. Take j_0 large enough such that $\sigma_u(z_j) \le K_u$ for all $j \ge j_0$, where $K_u > 0$ is as in condition U_3. Then $\{z_j\}_{j\ge j_0}$ is bounded and hence it has a weak accumulation point $\tilde{z}$. By (H0) we have that $\tilde{z} \in D^{-1}(x_{k_0})$. Because σ_u is weakly lower semicontinuous we also have $\sigma_u(\tilde{z}) \le \lim_{j\to\infty}\sigma_u(z_j) = 0$ so $\tilde{z} = 0 \in D^{-1}(x_{k_0})$. This yields x_{k_0} feasible for the problem and hence $h(x_{k_0}) \ge h(x^*)$, a contradiction with (5.8) for $k = k_0$. This implies that our claim is true and $a_k > 0$ for every k. By (5.8) we have

$$0 < a_k < h(x^*) - h(x_k) = \beta(0) - h(x_k).$$

By definition of a_k this implies the existence of $z_k \in D^{-1}(x_k)$ with $-p(0,-k\sigma_u(z_k)) < \beta(0) - h(x_k)$. Using (5.6) we can write

$$0 \le \psi(k\sigma_u(z_k)) \le -p(0,-k\sigma_u(z_k)) < \beta(0) - h(x_k) = h(x^*) - h(x_k). \tag{5.9}$$

Note that h is weakly lower semicontinuous and $\{x_k\}$ is bounded (because it converges to x^*). Therefore, there exists $L > 0$ such that $h(x_k) \geq -L$ for every k. Using also (5.9) we get

$$\psi(k\sigma_u(z_k)) \leq \beta(0) + L.$$

If (i) holds, the above implies that $\lim_{k\to\infty} \sigma_u(z_k) = 0$. On the other hand, if (ii) holds, the rightmost expression in (5.9) must tend to zero, and because $\psi(0) = 0$ we must have $\lim_{k\to\infty} \sigma_u(z_k) = 0$. Therefore, either under (i) or (ii) we have $\lim_{k\to\infty} \sigma_u(z_k) = 0$. By an argument similar to the one used above (which uses condition U_3) we conclude again that $\{z_k\}$ converges weakly to 0. Because $z_k \in D^{-1}(x_k)$ we get $\beta(z_k) = \inf_{x\in D(z_k)} h(x) \leq h(x_k)$ and (5.9) yields

$$\frac{\beta(z_k) - \beta(0)}{\sigma_u(z_k)} \leq \frac{h(x_k) - h(x^*)}{\sigma_u(z_k)} \leq \frac{p(0, -k\sigma_u(z_k))}{\sigma_u(z_k)} \leq \frac{-\psi(k\sigma_u(z_k))}{\sigma_u(z_k)}.$$

Using the above expression and the σ_u-calmness of β at 0 we get

$$\begin{aligned} -\infty < \liminf_k \tfrac{\beta(z_k)-\beta(0)}{\sigma_u(z_k)} &\leq \liminf_k \tfrac{p(0,-k\sigma_u(z_k))}{\sigma_u(z_k)} \leq \liminf_k \tfrac{-\psi(k\sigma_u(z_k))}{\sigma_u(z_k)} \\ &= \liminf_k (-k)\tfrac{\psi(k\sigma_u(z_k))}{(k\sigma_u(z_k))} \leq \liminf_k (-k)\delta = -\infty, \end{aligned}$$

where we used (H2) in the last inequality. The above expression entails a contradiction, and hence x^* must be a local minimizer of problem (5.3). ∎

6. Application to Lagrangian scheme. In this section we apply the abstract convexity results in order to analyze the properties of our augmented Lagrangian scheme. We start by defining exact penalty representation.

DEFINITION 6.1. *We say that* $(\bar{u}, \bar{y}) \in U \times Y$ supports an exact penalty representation *for problem* (P) *if there exists* $\bar{r} > 0$ *such that, for all* $r \geq \bar{r}$, *the original problem* (P) *is equivalent to minimizing* $l(\cdot, (r, \bar{y}, \bar{u}))$ *in the sense that*

$$\begin{aligned} &\text{(a)} \quad \beta(0) = \inf_{x\in X} l(x, (r, \bar{y}, \bar{u})), \ \textit{and} \\ &\text{(b)} \quad \operatorname{argmin}_{x\in X} h(x) = \operatorname{argmin}_{x\in X} l(x, (r, \bar{y}, \bar{u})) \end{aligned} \tag{6.1}$$

for all $r \geq \bar{r}$. *In this situation, the value* $\bar{r}$ *is said to be an exact* penalty parameter.

Remark 6.1. In the particular case in which the Lagrangian is the generalized penalty function (see Remark 3.5), the value of $\bar{r}$ reduces to the classical exact penalty parameter.

The previous definitions readily give the following characterization of (6.1)(a).

LEMMA 6.1. *With the notation of Definition* 6.1, *the following statements are equivalent:*

(i) $(\bar{u}, \bar{y})$ *verifies* (6.1)(a) *for the threshold* $r' > 0$.

(ii) $(r, \bar{y}, \bar{u}) \in \partial_\rho \beta(0)$ *for all* $r \geq r'$.

Proof. Using the definition of Lagrangian and perturbation function we can rewrite (6.1)(a) as

$$\begin{aligned} \beta(0) &= \inf_{x\in X} \inf_{z\in Z} f(x,z) - p(\nu_{\bar{y}}(z), -r\sigma_{\bar{u}}(z)) \\ &= \inf_{z\in Z} \inf_{x\in X} f(x,z) - p(\nu_{\bar{y}}(z), -r\sigma_{\bar{u}}(z)) \\ &= \inf_{z\in Z} \beta(z) - p(\nu_{\bar{y}}(z), -r\sigma_{\bar{u}}(z)) \\ &\leq \beta(z) - p(\nu_{\bar{y}}(z), -r\sigma_{\bar{u}}(z)) \end{aligned}$$

for all $z \in Z$ and all $r \geq r'$. Since $\sigma_{\bar{u}}(0) = \nu_{\bar{y}}(0) = 0$ this yields $(r, \bar{y}, \bar{u}) \in \partial_\rho \beta(0)$ for all $r \geq r'$. □

Remark 6.2. Proposition 4.3(b) provides conditions under which $\sigma_{\bar{u}}$-calmness at 0 implies that $\partial_\rho \beta(0) \neq \emptyset$. Applying also Lemma 6.1, we conclude that $\sigma_{\bar{u}}$-calmness at 0 implies the existence of $\bar{y} \in Y$ and $r' > 0$ such that $(\bar{u}, \bar{y})$ verifies (6.1)(a) for the threshold $r' > 0$.

Remark 6.3. Under the assumptions of Proposition 5.1, problem (P) has solutions. Indeed, because $f : X \times Z \to \bar{\mathbb{R}}$ is weakly lower semicontinuous and weakly level-compact, φ must be weakly lower semicontinuous and level-bounded, which clearly yields the nonemptiness and boundedness of the solution set.

Combining (6.1) with the definition of l we see that the exact penalty representation is connected with the optimal value and solution set of the family of problems given by

$$P(r, y, u) \qquad \inf_{(x,z) \in X \times Z} \{f(x, z) - p(\nu_{\bar{y}}(z), -r\sigma_{\bar{u}}(z))\}. \tag{6.2}$$

Fix $(u, y) \in U \times Y$ and consider all values of $r > 0$. The set

$$\cup_{r>0} \{(x_r, z_r) \mid (x_r, z_r) \text{ solves } P(r, y, u)\}$$

is called an *optimal path* for problem (P). Let us define the *primal optimal path* as the set

$$\cup_{r>0} \{x_r \mid (x_r, z_r) \text{ solves } P(r, y, u)\}.$$

Our next step is to establish conditions under which every weak accumulation point of the primal optimal path is a solution of the original problem.

THEOREM 6.1. *Assume all hypotheses of Theorem* 4.1 *hold and consider in* (6.2) *a duality parameterization $f : X \times Z \to \bar{\mathbb{R}}$ which is weakly lower semicontinuous and weakly level-compact. Then the following hold:*

(1) *There exists $r_0 > 0$ such that (x_r, z_r) solves problem $P(r, \bar{y}, \bar{u})$ for all $r > r_0$. In other words, the optimal path*

$$Q_{r_0} := \cup_{r \geq r_0} \{(x_r, z_r) \mid (x_r, z_r) \text{ solves } P(r, \bar{y}, \bar{u})\}$$

is well defined.

(2) *The sequence $\{z_r\}$ converges weakly to* 0*, and every weak accumulation point of $\{x_r\}$ is a solution of problem* (P).

Proof. Let us prove (1). Our assumptions on f and Proposition 5.1 imply that β is weakly lower semicontinuous. By Theorem 4.1(i) we have that for some $r_0 > 0$ and for every $r > r_0$ there exists $z_r \in Z$ such that $(r, \bar{y}, \bar{u}) \in \partial_\rho \beta(z_r)$. This means that for all $r > r_0$ we can write

$$\beta(z_r) - p(\nu_{\bar{y}}(z_r), -r\sigma_{\bar{u}}(z_r)) \leq \beta(z) - p(\nu_{\bar{y}}(z), -r\sigma_{\bar{u}}(z))$$

for all $z \in Z$. The above inequality implies that $z_r \in \operatorname{argmin} \{\beta(z) - p(\nu_{\bar{y}}(z), -r\sigma_{\bar{u}}(z))\}$ for all $r > r_0$, and taking $z = 0$ we get $\beta(z_r) \leq \beta(0)$. Fix $r > r_0$. By definition of β we can find a sequence $\{x_k\}$ such that

$$f(x_k, z_r) < \beta(z_r) + \frac{1}{k} \leq \beta(0) + 1.$$

By Theorem 4.1, $\{z_r\}$ converges weakly to 0. By the weak-level-compactness assumption on f we know that there exists a weak neighborhood W of 0 such that for all $z \in W$ we have

$$\{x \in X \,:\, f(x,z) \le \beta(0)+1\} \subset \tilde{B},$$

where $\tilde{B}$ is weakly compact. Without restriction, we can assume that $z_r \in W$ for all $r > r_0$. Therefore, $\{x_k\} \subset \tilde{B}$, and hence it has a weak accumulation point, which we call x_r. By weak lower semicontinuity of f, we can write

$$f(x_r,z_r) \le \liminf_k f(x_k,z_r) \le \liminf_k \beta(z_r) + \frac{1}{k} = \beta(z_r).$$

Hence $f(x_r,z_r) = \beta(z_r)$. Altogether,

$$\begin{aligned} &\inf_{x,z}\{f(x,z) - p(\nu_{\bar{y}}(z), -r\sigma_{\bar{u}}(z))\} = \inf_z\{\beta(z) - p(\nu_{\bar{y}}(z), -r\sigma_{\bar{u}}(z))\} \\ &\quad = \beta(z_r) - p(\nu_{\bar{y}}(z_r), -r\sigma_{\bar{u}}(z_r)) = f(x_r,z_r) - p(\nu_{\bar{y}}(z_r), -r\sigma_{\bar{u}}(z_r)), \end{aligned} \tag{6.3}$$

so (x_r, z_r) solves $P(r,\bar{y},\bar{u})$ for all $r > r_0$. This proves (1).

From (6.3) we have for every $x \in X$

$$f(x_r,z_r) - p(\nu_{\bar{y}}(z_r), -r\sigma_{\bar{u}}(z_r)) \le f(x,0). \tag{6.4}$$

The upper semicontinuity of p and the fact that $\{z_r\}$ converges weakly to 0 allow us to write

$$0 \le \liminf_{r\to\infty} -p(\nu_{\bar{y}}(z_r), -r\sigma_{\bar{u}}(z_r)).$$

Therefore there exists $r_1 \ge r_0$ such that

$$p(\nu_{\bar{y}}(z_r), -r\sigma_{\bar{u}}(z_r)) \le 1$$

for all $r > r_1$. Now fix $x \in X$. Using (6.4) and the above inequality for $r > r_1$ we get

$$f(x_r,z_r) \le f(x,0) + 1.$$

Using again the weak-level-compactness assumption we conclude that $\{x_r\}_{r>r_1}$ has weak accumulation points. Let us prove now that every weak accumulation point of $\{x_r\}$ is a solution of problem (P). Take x^* a weak accumulation point of $\{x_r\}$ and $\{x_{r_j}\}$ a subnet weakly converging to x^*. By (6.4) and weak lower semicontinuity,

$$f(x^*,0) \le \liminf_j f(x_{r_j},z_{r_j}) - p(\nu_{\bar{y}}(z_{r_j}), -r\sigma_{\bar{u}}(z_{r_j})) \le f(x,0),$$

where we also used the fact that $\{z_r\}$ converges weakly to 0. Now taking $\inf_{x\in X}$ we conclude that $f(x^*,0) \le \beta(0)$, which implies that x^* solves (P). This establishes (2). □

Remark 6.4. Assume there exists $(\bar{u},\bar{y})$ supporting an exact penalty representation for problem (P) with threshold $r' > 0$. Then by Lemma 6.1 we will have $(r,\bar{y},\bar{u}) \in \partial_\rho\beta(0)$ for all $r \ge r'$. In this case inequality (4.4) holds for all $z \in Z$. An application of Proposition 4.2 allows us to obtain a converse of this fact, where it is enough to check (4.4) in some neighborhood of 0. The result below, which is a criterion for the existence of exact penalty representation, has been established in

[13, Theorem 11.61] for a convex augmenting term (i.e., where $\sigma_u = \sigma$ for all $u \in U$ and σ is proper, lower semicontinuous, and convex) in finite dimensions. Theorem 6.2 extends analogous results recently proved in [23, 25, 8].

THEOREM 6.2 (criterion for exact penalty representation). *Assume all hypotheses of Theorem* 4.1 *hold (in particular,* $\operatorname{dom}\beta^\rho \neq \emptyset$*). Let* $(r, \bar{y}, \bar{u}) \in \operatorname{dom}\beta^\rho$ *be such that the function* $\sigma_{\bar{u}}$ *verifies condition* U_4*. Then* $(\bar{u}, \bar{y})$ *supports an exact penalty representation for problem* (P) *if there exists* $r' > 0$ *and a neighborhood* $V \subset Z$ *of* 0 *such that for all* $z \in V$ *the following inequality holds:*

$$\beta(0) \leq \beta(z) - p(\nu_{\bar{y}}(z), -r'\sigma_{\bar{u}}(z)). \tag{6.5}$$

Proof. We are in conditions of Proposition 4.2, which, together with (6.5), yields the existence of an $r^* > r'$ such that $(r, \bar{y}, \bar{u}) \in \partial_\rho\beta(0)$ for all $r \geq r^*$. Using now Lemma 6.1 we conclude that (6.1)(a) holds, or, equivalently,

$$\beta(0) \leq \beta(z) - p(\nu_{\bar{y}}(z), -r^*\sigma_{\bar{u}}(z)) \quad \text{for all } z \in Z. \tag{6.6}$$

An important consequence of the above inequality and condition U_4 is that

$$\operatorname{argmin}_{z\in Z}\beta(z) - p(\nu_{\bar{y}}(z), -r\sigma_{\bar{u}}(z)) = \{0\} \tag{6.7}$$

for all $r > r^*$. Indeed, fix $z' \neq 0$. Take $r > r^*$ and choose W a neighborhood of 0 such that $z' \notin W$. Call $c(W) := \inf_{z\notin W} \sigma_{\bar{u}}(z)$. We know by U_4 that $c(W) > 0$. We can write

$$\begin{aligned}
\beta(z') - p(\nu_{\bar{y}}(z'), -r\sigma_{\bar{u}}(z')) &= \beta(z') - p(\nu_{\bar{y}}(z'), -r^*\sigma_{\bar{u}}(z')) \\
&\quad + p(\nu_{\bar{y}}(z'), -r^*\sigma_{\bar{u}}(z')) - p(\nu_{\bar{y}}(z'), -r\sigma_{\bar{u}}(z')) \\
&\geq \beta(0) + \psi((r - r^*)\sigma_{\bar{u}}(z')) \\
&\geq \beta(0) + \psi((r - r^*)\sigma_{\bar{u}}(z')) > \beta(0),
\end{aligned}$$

where we used (6.6) and assumption (b) on p. Now we proceed to establish (6.1)(b), i.e., that there exists $r_1 \geq r^*$ such that for all $r > r_1$ we have

$$\operatorname{argmin}_{x\in X}\varphi(x) = \operatorname{argmin}_{x\in X} l(x, (r, \bar{y}, \bar{u})). \tag{6.8}$$

Define the function $h(x, z) := f(x, z) - p(\nu_{\bar{y}}(z), -r\sigma_{\bar{u}}(z))$. It is directly checked from the definitions that

$$\begin{aligned}
&\text{(a)} \quad \operatorname{argmin}_{(x,z)\in X\times Z} h(x, z) = \{(x', z') \mid x' \in \operatorname{argmin}_{x\in X} h(x, z') \text{ and} \\
&\qquad\qquad z' \in \operatorname{argmin}_{z\in Z}\beta(z) - p(\nu_{\bar{y}}(z), -r\sigma_{\bar{u}}(z))\} \\
&\text{(b)} \qquad\qquad = \{(x', z') \mid x' \in \operatorname{argmin}_{x\in X} l(x, (r, \bar{y}, \bar{u})) \text{ and} \\
&\qquad\qquad z' \in \operatorname{argmin}_{z\in Z} h(x', z)\}.
\end{aligned} \tag{6.9}$$

Let us start the proof of (6.8) by taking $\bar{x} \in \operatorname{argmin}_{x\in X} l(x, (\bar{u}, \bar{y}, r))$. We claim that there exists $r_1 > r^*$ such that for all $r > r_1$ we have

$$\operatorname{argmin}_{z\in Z} h(\bar{x}, z) = \{0\}. \tag{6.10}$$

By the definition of Lagrangian, the definition of $\bar{x}$, and (6.6) we can write

$$\begin{aligned}
\inf_{z\in Z} h(\bar{x}, z) &= l(\bar{x}, (r, \bar{y}, \bar{u})) = \inf_{x\in X} l(x, (r, \bar{y}, \bar{u})) \\
&= \inf_{x\in X} \inf_{z\in Z} h(x, z) = \inf_{z\in Z} \inf_{x\in X} h(x, z) \\
&= \inf_{z\in Z} \beta(z) - p(\nu_{\bar{y}}(z), -r\sigma_{\bar{u}}(z)) = \beta(0).
\end{aligned} \tag{6.11}$$

Inequality (6.6) yields $\inf_{z\in Z} f(x,z) - p(\nu_{\bar y}(z), -r^*\sigma_{\bar u}(z)) \geq \inf_{z\in Z}\beta(z) - p(\nu_{\bar y}(z), -r^*\sigma_{\bar u}(z)) \geq \beta(0) > -\infty$. Thus the optimal value of $P(r^*, \bar y, \bar u)$ is greater than $-\infty$. Fix $r > r^*$. Using (6.11) and an argument similar to that in the proof of Theorem 4.1, we can find a sequence $\{z_n\} \subset Z$ such that $\lim_{n\to\infty}\beta(z_n) - p(\nu_{\bar y}(z_n), -r\sigma_{\bar u}(z_n)) = \beta(0)$ and

$$\sigma_{\bar u}(z_n) \leq \frac{\psi^{-1}(\beta(0) + 1 - m(r^*))}{r - r^*}$$

for all $n \geq n_0$. Take now $r_1 > r^* + \frac{\psi^{-1}(\beta(0)+1-m(r^*))}{K_{\bar u}}$, where $K_{\bar u}$ is as in U_3. This yields $\sigma_{\bar u}(z_n) \leq K_{\bar u}$, and hence by U_3 the sequence $\{z_n\} \subset Z$ is bounded. Take a weak accumulation point z_r along with a subsequence $\{z_{n_j}\} \subset \{z_n\}$ weakly converging to it. Then by lower semicontinuity of the functions and (6.11), we get

$$f(\bar x, z_r) - p(\nu_{\bar y}(z_r), -r\sigma_{\bar u}(z_r)) \leq \liminf_{j\to\infty} f(\bar x, z_{n_j}) - p(\nu_{\bar y}(z_{n_j}), -r\sigma_{\bar u}(z_{n_j})) = \beta(0).$$

By definition of β we always have $f(\bar x, z_r) \geq \beta(z_r)$. Combining the last two facts and (6.6) we get

$$\beta(0) \leq \beta(z_r) - p(\nu_{\bar y}(z_r), -r\sigma_{\bar u}(z_r)) \leq \beta(0),$$

so by (6.11) we must have $z_r \in \operatorname{argmin}_{z\in Z}\beta(z) - p(\nu_{\bar y}(z), -r\sigma_{\bar u}(z))$ for all $r > r_1$ and by (6.7) this means that $z_r = 0$. This establishes (6.10). As a consequence of (6.10), we have that the pair $(\bar x, 0) \in X \times Z$ verifies the conditions $\bar x \in \operatorname{argmin}_{x\in X} l(x, (r, \bar y, \bar u))$ and $0 \in \operatorname{argmin}_{z\in Z} h(\bar x, z)$. Now using (6.9)(b) we conclude that $(\bar x, 0) \in \operatorname{argmin}_{(x,z)\in X\times Z} h(x,z)$. Using the latter fact in (6.11) we obtain

$$\inf_{z\in Z} h(\bar x, z) = \inf_{x\in X}\inf_{z\in Z} h(x,z) = h(\bar x, 0) = f(\bar x, 0).$$

Hence, for all $x \in X$ we have that

$$\varphi(\bar x) = f(\bar x, 0) = h(\bar x, 0) \leq h(x,0) = f(x,0) = \varphi(x),$$

which yields $\bar x \in \operatorname{argmin}_{x\in X}\varphi(x)$.

Conversely, now take $\bar x \in \operatorname{argmin}_{x\in X}\varphi(x)$ and $r > r_1$. For all $x \in X$ we have that

$$h(\bar x, 0) = \varphi(\bar x) \leq \varphi(x) = f(x,0) = h(x,0),$$

so that $\bar x \in \inf_{x\in X} h(x,0)$. Since $r_1 > r^*$ we can use (6.7), and hence we have that $0 \in \operatorname{argmin}_{z\in Z}\beta(z) - p(\nu_{\bar y}(z), -r\sigma_{\bar u}(z))$, so the pair $(\bar x, 0)$ belongs to the set on the right-hand side of (6.9)(a). We then conclude again that $(\bar x, 0) \in \operatorname{argmin}_{(x,z)\in X\times Z} h(x,z)$. Use now (6.9)(b) to get $\bar x \in \operatorname{argmin}_{x\in X} l(x, (r, \bar y, \bar u))$, as we wanted.

Hence we conclude that $\operatorname{argmin}_{x\in X}\varphi(x) = \operatorname{argmin}_{x\in X} l(x, (r, \bar y, \bar u))$ for all $r > r_1$, where $r_1 > r^*$. □

7. Semi-infinite programming. Consider semi-infinite stochastic programming problems of the form

$$\text{minimize} \quad h(x) \quad \text{subject to} \quad g(x,\xi) \leq 0 \text{ a.e. } \xi \in \Xi, \tag{7.1}$$

where $h : \mathbb{R}^n \to \mathbb{R}_{+\infty}$ is a lower semicontinuous function, and Ξ is a set equipped with a σ-algebra $\mathcal{M}$ and a finite measure μ defined on $\mathcal{M}$. In the above formulation,

a feasible point x verifies the inequalities $g(x,\xi) \le 0$ a.e. $\xi \in \Xi$ if and only if there is a subset $A \in \mathcal{M}$ such that $\mu(A) = 0$ and the inequality $g(x,\xi) \le 0$ holds for every $\xi \in \Xi \setminus A$. This kind of formulation is relevant, for example, in stochastic programming (cf. [14, 15, 16]). Semi-infinite programming problems are studied in detail in [1, 7, 22, 20]. We assume that

(H4) h is coercive, i.e., $\lim_{\|x\|\to\infty} h(x) = +\infty$;

(G1) for every fixed $x \in \mathbb{R}^n$ the function $g(x,\cdot) : \Xi \to \mathbb{R}$ is measurable;

(G2) $g(\cdot,\xi) : \mathbb{R}^n \to \mathbb{R}$ is lower semicontinuous in $\mathbb{R}^n$.

Call $C := \{x \in \mathbb{R}^n \mid g(x,\xi) \le 0 \text{ a.e. } \xi \in \Xi\}$ the feasible set of problem (7.1). Denote by h^* the optimal value of problem (7.1). Let $X = \mathbb{R}^n$ and $Z = L_p$. Fix a function $\eta : L_p \to L_p$ such that

(A_0) the function $\eta(\cdot)(\xi) : L_p \to \mathbb{R}$ is weakly lower semicontinuous a.e. in Ξ and $\eta(0) = 0$.

Define the point-to-set mapping $D : L_p \rightrightarrows X$ as

$$D(z) := \{x \in \mathbb{R}^n \mid g(x,\xi) + \eta(z)(\xi) \le 0 \text{ a.e. } \xi \in \Xi\}. \tag{7.2}$$

Note that the assumption $\eta(0) = 0$ yields $D(0) = C$. We can formulate problem (7.1) in terms of a dualizing parameterization f which verifies the assumptions of Theorem 4.1. Indeed, let

$$\varphi(x) := h(x) + \delta_C(x) \quad \text{and} \quad f(x,z) := h(x) + \delta_{D(z)}(x).$$

Then it is direct to check that f is a dualizing parametrization for φ.

LEMMA 7.1. *If h is lower semicontinuous and assumptions* (H4), (G1), (G2), *and* (A_0) *hold, then f is weakly level-compact and weakly lower semicontinuous. Consequently, the perturbation function $\beta : L_p \to \mathbb{R} \cup \{-\infty\}$ is weakly lower semicontinuous.*

Proof. By Proposition 5.2 and (H4), it is enough to check assumptions (H0) and (H1). Condition (H1) is the lower semicontinuity of h. Condition (H0) follows from conditions (G2) and (A0). ∎

Remark 7.1. Let us note that the coercivity assumption (H4) and lower semicontinuity of h directly yield nonemptiness and boundedness of the solution set.

Let us now construct the Lagrangian scheme. Let Y and U be two sets and define $\Omega := \mathbb{R}_+ \times Y \times U$. We use $p(a,b) = a + b$ in (3.3) to get $\rho(z,(r,y,u)) = \nu_y(z) - r\sigma_u(z)$, where the families $\{\nu_y\}_{y\in Y}$ and $\{\sigma_u\}_{u\in U}$ are functions from Z to $\mathbb{R}$ verifying properties Y_1–Y_2 and U_1–U_4, respectively (for instance, as in Remark 3.4, we can take $Y = L_q(\Xi,\mathcal{M},\mu)$ with $1/p + 1/q = 1$ and $\nu_y(z) = \langle y, z\rangle$; and we recover a classical augmented Lagrangian). The definition of f yields

$$\beta(z) = \inf_{x\in X} f(x,z) = \inf_{x\in D(z)} h(x).$$

Recall that $x \in D(z)$ if and only if $z \in D^{-1}(x) := \{z' \mid x \in D(z')\}$. Then the augmented Lagrangian corresponding to ρ in a fixed element $(x,(r,y,u))$ has the form

$$\begin{aligned} l(x,(r,y,u)) &= \inf_{z\in Z}\{f(x,z) - \rho(z,\omega)\} = \inf_{z\in D^{-1}(x)} h(x) - \nu_y(z) + r\sigma_u(z) \\ &= h(x) + \inf_{z\in D^{-1}(x)}\{-\nu_y(z) + r\sigma_u(z)\}. \end{aligned} \tag{7.3}$$

From the expression for the Lagrangian we obtain the family of problems $P(r,y,u)$ given by

$$\min_{(x,z)\in\mathbb{R}^n\times L_p} f(x,z) - \nu_y(z) + r\sigma_u(z). \tag{7.4}$$

The optimal path associated with a fixed pair $(\overline{y}, \bar{u}) \in Y \times U$ is defined as the set

$$\cup_{r>0}\{(x_r, z_r) \mid (x_r, z_r) \text{ solves } P(r, \overline{y}, \bar{u})\}.$$

We state below conditions guaranteeing existence of the optimal path for r large enough.

THEOREM 7.1. *Under the assumptions of Lemma* 7.1 *and Theorem* 4.1 *we have that*

(a) *there exist* $r_0 > 0$ *and* $\bar{y} \in Y$ *such that* (x_r, z_r) *solves* $P(r, \overline{y}, \bar{u})$ *for all* $r > r_0$*;*
(b) *the sequence* $\{z_r\}$ *converges weakly to* 0*, and every accumulation point of* $\{x_r\}$ *is a solution of problem* (7.1).

Proof. Note that the assumptions of Theorem 6.1 hold, and the conclusions of the latter theorem are precisely conclusions (a) and (b). ∎

Next we apply Proposition 5.3 in order to prove that calmness of the perturbation function implies the existence of an exact penalty parameter.

PROPOSITION 7.1. *Consider the Lagrangian scheme defined in* (7.3)–(7.4). *Assume that* h *is lower semicontinuous and suppose also that* (H4), (G2), *and* (A_0) *hold. Let* σ_u *satisfy* U_1–U_3 *and let* Y *be such that* $0 \in Y$ *and* $\nu_0(z) = 0$ *for every* $z \in Z$. *If* β *is* σ_u*-calm at* 0*, then for every solution* x^* *of problem* (7.1)*, there exists* $M > 0$ *such that* x^* *is a local solution of*

$$\min_{x \in X} \tilde{l}(x, (M, u)),$$

where $\tilde{l}(x, (M, u)) := l(x, (M, 0, u))$.

Proof. By Proposition 5.3, it is enough to check that assumptions (H0)–(H3) hold for our Lagrangian scheme. Note first that our choice $p(a, b) = a + b$ entails $\psi(t) = t$ for all $t \geq 0$, yielding (H2)–(H3). As noted before, (H0) is a consequence of (G2) and (A_0). Finally, (H1) is our basic assumption of lower semicontinuity of h. ∎

The following result is a direct consequence of Theorem 6.2.

THEOREM 7.2. *Assume all hypotheses of Theorem* 4.1 *hold, where* $(\bar{r}, \overline{y}, \bar{u}) \in \operatorname{dom} \beta^\rho$. *Then* $(\bar{r}, \overline{y}, \bar{u})$ *supports an exact penalty representation for problem* (7.1) *if there exist* $r' > 0$ *and a neighborhood* V *of* 0 *in* L_p *such that for all* $z \in V$ *we have*

$$h^* \leq [\inf_{x \in D(z)} h(x)] - \nu_{\overline{y}}(z) + r' \sigma_{\overline{u}}(z).$$

Combining the above result with Example 4.1, we can characterize nonemptiness $\partial_\rho \beta(0)$. The proof of this result follows readily from Corollary 4.1 and the fact that p and the functions $\nu_{\overline{y}}$ and $\sigma_{\overline{u}}$ verify conditions (i) and (ii) of Definition 4.2.

COROLLARY 7.1. *Let the families of functions* $\{\nu_y, \sigma_u\}$ *be given as in Example* 4.1. *Assume all hypotheses of Theorem* 4.1 *hold, where* $(\bar{r}, \overline{y}, \bar{u}) \in \operatorname{dom} \beta^\rho$. *Then* $\partial_\rho \beta(0) \neq \emptyset$ *if and only if the perturbation function* β *is* $\sigma_{\overline{u}}$*-calm at* 0.

Remark 7.2. We apply the result above to the augmented Lagrangians constructed in [13, Chapter 11], where the family $\{\nu_y\}$ is the dual of Z (i.e., $\nu_y(z) = \langle y, z \rangle$ for all $z \in Z$) and the family $\{\sigma_u\}$ has a single element σ with $\sigma(0) = 0$. The function p is given by $p(a, b) = a + b$ so that the coupling function becomes $\rho(z, \omega) = \rho(z, (r, y)) = \langle y, z \rangle - r\sigma(z)$. We have that ν_y, p, and σ verify conditions (i) and (ii) of Definition 4.2 when $\limsup_{z \to 0, z \neq 0} \frac{\|z\|}{\sigma(z)} < +\infty$. When $\sigma(\cdot) := \|\cdot\|$ then the latter condition trivially holds.

An important particular case of problem (7.1) is the semi-infinite programming problem, stated as

$$\text{minimize} \quad h(x) \quad \text{subject to} \quad g(x, t) \leq 0,\ t \in T, \tag{7.5}$$

where $T \subset \mathbb{R}^p$ is a compact set, $h : \mathbb{R}^n \to \mathbb{R}_{+\infty}$ is lower semicontinuous, and $g(\cdot, t) : \mathbb{R}^n \to \mathbb{R}$, $g(x, \cdot) : \mathbb{R}^p \to \mathbb{R}$ are also lower semicontinuous for every $(x, t) \in \mathbb{R}^n \times \mathbb{R}^p$. A popular choice of penalty function for these problems is the standard L_∞-penalty function (see, e.g., [4, 5]), given by

$$\tilde{l}(x, r) := h(x) + r \max_{t \in T} |[g(x, t)]_+|,$$

where $[a]_+ = \max\{a, 0\}$ for every $a \in \mathbb{R}$. In the following example, we show that our setting includes this kind of penalty function as a particular case.

Example 7.1. Consider the semi-infinite programming problem (7.5) and take $Z := L_\infty(T)$ with the norm $\|z\|_\infty := \sup_{t \in T} |z(t)|$. Assume also that $g(x, \cdot) \in L_\infty(T)$ for every $x \in \mathbb{R}^n$ (in other words, $g(x, \cdot)$ is bounded above on the set T for every x). Let $\eta : L_\infty(T) \to L_\infty(T)$ be defined as $\eta(z) = z$. Our perturbed problems become

$$\text{minimize} \quad h(x) \quad \text{subject to} \quad g(x, t) + z(t) \leq 0,\ t \in T.$$

As in (7.2), consider the set

$$D(z) := \{x \in \mathbb{R}^n \ : \ g(x, t) + z(t) \leq 0,\ t \in T\}$$

and the duality parameterization $f(x, z) := h(x) + \delta_{D(z)}(x)$. If we take $\nu_y = 0$ for every $y \in Y$ and $\sigma_u = \|\cdot\|_\infty$ for every $u \in U$, then (7.3) gives

$$\tilde{l}(x, r) = h(x) + r \inf_{z \in D^{-1}(x)} \|z\|_\infty.$$

We claim that $\inf_{z \in D^{-1}(x)} \|z\|_\infty = \| [g(x, \cdot)]_+ \|_\infty$. Indeed, define $\bar{z}(t) := \min\{-g(x, t), 0\}$ for all $t \in T$. It can be checked that (i) $\bar{z} \in D^{-1}(x)$, (ii) $\|\bar{z}\|_\infty = \| [g(x, \cdot)]_+ \|_\infty$, and (iii) $\|z\|_\infty \geq \|\bar{z}\|_\infty$ for every $z \in D^{-1}(x)$. Items (i) and (ii) follow directly from the definitions, so let us prove (iii). Assume there exist $z \in D^{-1}(x)$ and $a \in \mathbb{R}$ with $\|z\|_\infty < a < \|\bar{z}\|_\infty = \| [g(x, \cdot)]_+ \|_\infty$. This implies that there exists $\bar{t} \in T$ such that

$$|z(t)| < a < [g(x, \bar{t})]_+ \tag{7.6}$$

for every $t \in T$. The above inequality can hold only if $g(x, \bar{t}) > 0$. Now using the fact that $z \in D^{-1}(x)$ we have that $z(\bar{t}) \leq -g(x, \bar{t}) < 0$. By (7.6) for $t = \bar{t}$ we get $-z(\bar{t}) = |z(\bar{t})| < a < [g(x, \bar{t})]_+ = g(x, \bar{t})$, which yields $z(\bar{t}) + g(x, \bar{t}) > 0$, a contradiction to the fact that $z \in D^{-1}(x)$. Hence the claim is true and the Lagrangian simplifies to

$$\tilde{l}(x, r) = h(x) + r \max_{t \in T} |[g(x, t)]_+|,$$

which is the classical L_∞-penalty function. We point out that a similar analysis to the one in Lemma 7.1 can be carried out in order to prove lower semicontinuity of the perturbation function, so that the subsequent theoretical results, such as well-definedness of the central path and the criterion for exact penalty representation, can also be established for this problem with this kind of augmenting term.

Acknowledgment. The authors are very grateful to the two referees for their helpful and valuable suggestions, which greatly improved earlier versions of the manuscript.

REFERENCES

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

[2] A. Brøndsted and R. T. Rockafellar, *On the subdifferentiability of convex functions*, Proc. Amer. Math. Soc., 16 (1965), pp. 605–611.

[3] F. H. Clarke, *Optimization and Nonsmooth Analysis*, SIAM, Philadelphia, 1990.

[4] I. D. Coope and C. J. Price, *Exact penalty function methods for nonlinear semi-infinite programming*, in Semi-infinite Programming, Nonconvex Optim. Appl. 25, Kluwer Academic, Boston, MA, 1998, pp. 137–157.

[5] I. D. Coope and C. J. Price, *The L_∞ exact penalty function in semi-infinite programming*, in Computational Techniques and Applications: CTAC-89 (Brisbane, 1989), Hemisphere, New York, 1990, pp. 657–664.

[6] R. Gasimov and A. M. Rubinov, *Augmented Lagrangians for optimization problems with a single constraint*, J. Global Optim., 28 (2004), pp. 153–173.

[7] M. A. Goberna and M. A. López, *Semi-infinite Programming—Recent Advances*, Kluwer Academic, Dordrecht, The Netherlands, 2001.

[8] X. X. Huang and X. Q. Yang, *A unified augmented Lagrangian approach to duality and exact penalization*, Math. Oper. Res., 28 (2003), pp. 533–552.

[9] J. V. Outrata and J. Jarusek, *Duality Theory in Mathematical Programming and Optimal Control*, Kibernetika (Prague) Suppl. 20/21, Academy of Sciences of the Czech Republic, Prague, 1984/1985.

[10] J. P. Penot, *Augmented Lagrangians, duality and growth conditions*, J. Nonlinear Convex Anal., 3 (2002), pp. 283–302.

[11] R. T. Rockafellar, *Conjugate Duality and Optimization*, SIAM, Philadelphia, 1974.

[12] R. T. Rockafellar, *Lagrange multipliers and optimality*, SIAM Rev., 35 (1993), pp. 183–238.

[13] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer, Berlin, 1998.

[14] R. T. Rockafellar and R. J.-B. Wets, *Stochastic convex programming: Basic duality*, Pacific J. Math., 62 (1976), pp. 173–195.

[15] R. T. Rockafellar and R. J.-B. Wets, *Stochastic convex programming: Singular multipliers and extended duality*, Pacific J. Math., 62 (1976), pp. 507–522.

[16] R. T. Rockafellar and R. J.-B. Wets, *Stochastic convex programming: Relatively complete recourse and induced feasibility*, SIAM J. Control Optim., 14 (1976), pp. 574–589.

[17] A. M. Rubinov, X. X. Huang, and X. Q. Yang, *The zero duality gap property and lower semicontinuity of the perturbation function*, Math. Oper. Res., 27 (2002), pp. 775–791.

[18] A. M. Rubinov, *Abstract Convexity and Global Optimization*, Kluwer Academic, Dordrecht, The Netherlands, 2000.

[19] A. M. Rubinov and X. Q. Yang, *Lagrange-Type Functions in Constrained Non-convex Optimization*, Kluwer Academic, Dordrecht, The Netherlands, 2003.

[20] A. Shapiro, *On duality theory of convex semi-infinite programming*, Optimization, 54 (2005), pp. 535–543.

[21] I. Singer, *Abstract Convex Analysis*, Canadian Mathematical Society Series of Monographs and Advanced Texts, John Wiley, New York, 1997.

[22] O. Stein, *Bi-level Strategies in Semi-infinite Programming*, Kluwer Academic, Boston, MA, 2003.

[23] Y. Y. Zhou and X. Q. Yang, *Some results about duality and exact penalization*, J. Global Optim., 29 (2004), pp. 497–509.

[24] Y. Y. Zhou and X. Q. Yang, *Duality and penalization in optimization via an augmented Lagrangian function with applications*, J. Optim. Theory Appl., to appear.

[25] Y. Y. Zhou and X. Q. Yang, *Augmented Lagrangian function, non-quadratic growth condition and exact penalization*, Oper. Res. Lett., 34 (2006), pp. 127–134.

SIAM J. OPTIM.
Vol. 18, No. 2, pp. 437–460

METRIC SUBREGULARITY AND CONSTRAINT QUALIFICATIONS FOR CONVEX GENERALIZED EQUATIONS IN BANACH SPACES*

XI YIN ZHENG† AND KUNG FU NG‡

Abstract. Several notions of constraint qualifications are generalized from the setting of convex inequality systems to that of convex generalized equations. This is done and investigated in terms of the coderivatives and the normal cones, and thereby we provide some characterizations for convex generalized equations to have the metric subregularity. As applications, we establish formulas of the modulus of calmness and provide several characterizations of the calmness. Extending the classical concept of extreme boundary, we introduce a notion of recession cores of closed convex sets. Using this concept, we establish global metric subregularity (i.e., error bound) results for generalized equations.

Key words. metric subregularity, calmness, constraint qualification, normal cone, coderivative, recession core, generalized extreme point

AMS subject classifications. 90C31, 90C25, 49J52, 46B20

DOI. 10.1137/050648079

1. Introduction.

Let X and Y be Banach spaces and $F : X \to 2^Y$ a closed multifunction. Following Dontchev and Rockafellar [7], the multifunction F is said to be metrically subregular at a for $b \in F(a)$ if there exists $\tau \in [0, +\infty)$ such that

$$d(x, F^{-1}(b)) \le \tau d(b, F(x)) \quad \forall x \text{ close } to\ a. \tag{1.1}$$

The metric subregularity has already been studied by many authors under various names (cf. [2, 15, 21, 26, 32]).

Let A be a closed subset of X and b be a given point in Y. Consider the generalized equation with constraint

$$b \in F(x) \ \text{ subject to } x \in A, \tag{GEC}$$

which includes most of the systems in optimization. Let S denote the solution set of (GEC), that is, $S = \{x \in A : \ b \in F(x)\}$.

We say that (GEC) is metrically subregular at $a \in S$ if there exists $\tau \in (0, \infty)$ such that

$$d(x, S) \le \tau(d(b, F(x)) + d(x, A)) \quad \forall x \text{ close to } a. \tag{1.2}$$

When $F(x) = [f(x), +\infty)$, $b = 0$, and $A = X$, (GEC) reduces to the inequality system $f(x) \le 0$ and (1.2) means that this inequality has a local error bound at a.

*Received by the editors December 20, 2005; accepted for publication (in revised form) December 5, 2006; published electronically May 16, 2007. This research was supported by a postdoctoral fellowship scheme and a direct grant (CUHK) and an Earmarked Grant from the Research Grant Council of Hong Kong.

http://www.siam.org/journals/siopt/18-2/64807.html

†Department of Mathematics, Yunnan University, Kunming 650091, People's Republic of China (xyzheng@ynu.edu.cn). This author's work was also supported by the National Natural Science Foundation of People's Republic of China (grant 10361008) and the Natural Science Foundation of Yunnan Province, China (grant 2003A0002M).

‡Department of Mathematics, The Chinese University of Hong Kong, Shatin, New Territory, Hong Kong (kfng@math.cuhk.edu.hk).

Another special case is when $F(x) = [f(x), +\infty)$ and $b = \inf\{f(x) : x \in A\}$. In this case, (GEC) reduces to the following optimization problem:

$$\text{(OP)} \qquad \min f(x) \ \text{ subject to } x \in A$$

and (1.2) means that a is a weak sharp minimum of (OP). Error bounds and weak sharp minima have important applications in mathematical programming and have been extensively studied (cf. [3, 18, 19, 21, 34, 35]). In this paper, we mainly study the metric subregularity of (GEC) in the case when F and A are convex.

The notion of the basic constraint qualification (BCQ) of systems of continuous convex inequalities plays an important role in convex optimization and has been studied by many researchers (see, e.g., [11, 307–309 and 19, 20]). Dropping the continuity assumption and adopting the singular subdifferential, the authors [36] introduced and discussed the generalized BCQ and strong BCQ. Very recently, Hu [12] further studied the generalized BCQ and strong BCQ. In section 3, in terms of the coderivative, we extend the concept of the generalized BCQ and strong BCQ to cover the case of a generalized equation with constraint (GEC). Using the BCQ and strong BCQ, we provide several characterizations of the metric subregularity of (GEC).

A stronger condition is the metric regularity of a multifunction that has been well studied in variational analysis (see [13, 15, 22, 24, 30] and references therein). Explicitly, F is metrically regular at a for $b \in F(a)$ if there exists $\tau \in (0, +\infty)$ such that

$$d(x, F^{-1}(y)) \le \tau d(y, F(x)) \quad \forall (x, y) \text{ close } to\ (a, b). \tag{1.3}$$

It is well known (as the Robinson–Ursescu theorem) that (1.3) holds if F is a closed convex multifunction and $b \in \operatorname{int}(F(X))$. Under the assumption that both X and Y are finite dimensional, Mordukhovich [22] proved that F is metrically regular at a for $b \in F(a)$ if and only if $D^*F(a, b)^{-1}(0) = \{0\}$; moreover,

$$\inf\{\tau : \text{ (1.3) holds}\} = \|D^*F(a, b)^{-1}\|^+ = \limsup_{(x,y) \overset{\mathrm{Gr}(F)}{\longrightarrow} (a,b)} \|D^*F(x, y)^{-1}\|^+, \tag{1.4}$$

where $D^*F(a, b)$ is the coderivative of F at (a, b) and

$$\|D^*F(a, b)^{-1}\|^+ = \sup_{x^* \in B_{X^*}} \sup_{y^* \in D^*F(a,b)^{-1}(x^*)} \|y^*\|.$$

When X and Y are infinite dimensional, some sufficient and necessary conditions for the metric regularity were also established (see [17, 23], Mordukhovich's recent books [24, 25], and references therein). To the best of our knowledge, no one has considered duality formulas similar to (1.4) for the modulus of the metric subregularity. In section 3, we provide such formulas under the convexity assumption.

Similar to the relationship between Aubin's pseudo-Lipschitz property and the metric regularity, the calmness is related very closely to the metric subregularity. In section 4, as applications of results obtained in section 3, we consider the calmness of convex multifunctions. We establish formulas of the modulus of the calmness and present several characterizations of the calmness in terms of the normal cone and the coderivative. In this section, we also provide characterizations of the strong calmness. Reducing to special kinds of convex multifunctions such as that recently considered by Henrion and Jourani in [8], our approach sheds light on some existing results on

the calmness; in fact Corollary 4.2 provides a version that is sharper than the main result in [8].

The notion of the extreme point set of a convex set is very useful in convex analysis. In section 5, as an extension of an extreme point set, we introduce and discuss the notion of a recession core. In terms of recession cores, we study the global metric subregularity. In particular, we show that (GEC) is globally metrically subregular if and only if (GEC) has the τ-strong BCQ at each point of some recession core of the solution set S for some $\tau \in (0, +\infty)$. When the solution set S is a polyhedron, we obtain a sharp result that (GEC) is globally metrically subregular if and only if (GEC) has the BCQ at each point of some recession core of S. This implies in particular that if the graph of F is a polyhedron, then (GEC) is always globally metrically subregular; thus the classical Hoffman result on the error bound for linear inequality systems are extended and improved to cover some nonlinear inequality systems without the Slater condition.

2. Preliminaries. Throughout this paper, we assume that X and Y are Banach spaces. We denote by B_X and B_Y the closed unit balls of X and Y, respectively. For $a \in X$ and $\delta > 0$, let $B(a, \delta)$ denote the open ball with center a and radius δ.

For a closed convex subset A of X and $a \in A$, we use $T(A, a)$ to denote the tangent cone of A at a in Bouligand's sense. Thus $v \in T(A, a)$ if and only if there exist a sequence $\{a_n\}$ in A and a sequence $\{t_n\}$ of positive numbers convergent to 0 such that $\frac{a_n - a}{t_n}$ converges to v.

We denote by $N(A, a)$ the normal cone of A at a, that is,

$$N(A, a) := \{x^* \in X^* : \langle x^*, x - a\rangle \le 0 \ \ \forall x \in A\}.$$

Let $F : X \to 2^Y$ be a multifunction and denote by $\mathrm{Gr}(F)$ the graph of F, that is,

$$\mathrm{Gr}(F) := \{(x, y) \in X \times Y : \ y \in F(x)\}.$$

As usual, F is said to be closed (resp., convex) if $\mathrm{Gr}(F)$ is a closed (resp., convex) subset of $X \times Y$. It is known that F is convex if and only if

$$tF(x_1) + (1 - t)F(x_2) \subset F(tx_1 + (1 - t)x_2) \quad \forall x_1, x_2 \in X \text{ and } \forall t \in [0,\ 1].$$

For a closed convex multifunction F and $(x, y) \in \mathrm{Gr}(F)$, the tangent derivative $DF(x, y)$ of F at (x, y) is defined by

$$DF(x, y)(u) = \{v \in Y : \ (u, v) \in T(\mathrm{Gr}(F), (x, y))\} \quad \forall u \in X \tag{2.1}$$

(cf. [1]).

Let $D^*F(x, y)$ denote the coderivative of F at (x, y), which is defined by

$$D^*F(x, y)(y^*) := \{x^* \in X^* : \ (x^*, -y^*) \in N(\mathrm{Gr}(F), (x, y))\} \quad \forall y^* \in Y^* \tag{2.2}$$

(cf. [22, 26, 32]).

Let $G : X \to 2^Y$ be a sublinear multifunction (i.e., $\mathrm{Gr}(G)$ is a convex cone in $X \times Y$). As in Dontchev, Lewis, and Rockafellar [6], the outer norm and inner norm of G are, respectively, defined as

$$\|G\|^+ = \sup_{x \in B_X} \sup_{y \in Gx} \|y\| \quad \text{and} \quad \|G\|^- := \sup_{x \in B_X} \inf_{y \in Gx} \|y\|,$$

where the infimum and supremum over an empty set are understood as $+\infty$ and 0, respectively. For a convex cone C in X, let $\|G|_C\|^+$ and $\|G|_C\|^-$ be, respectively, defined by

$$\|G|_C\|^+ := \sup_{x\in B_X\cap C}\sup_{y\in Gx}\|y\| \quad\text{and}\quad \|G|_C\|^- := \sup_{x\in B_X\cap C}\inf_{y\in Gx}\|y\|. \tag{2.3}$$

We denote by $\mathrm{bd}(A)$ the topological boundary of a subset A of X. The following lemma is known (cf. [27, Proposition 1.3] or [28, Lemma 2.1]) and useful for us.

LEMMA 2.1. *Let X be a Banach space and A a closed convex nonempty subset of X. Then, for any $\beta \in (0,\ 1)$ and any $x \in X \setminus A$ there exist $z \in \mathrm{bd}(A)$ and $x^* \in N(A, z)$ with $\|x^*\| = 1$ such that*

$$\beta\|x - z\| < d(x, A) \ \text{ and } \ \beta\|x - z\| < \langle x^*, x - z\rangle.$$

3. BCQ, strong BCQ, and metric subregularity. Throughout this section, we assume that $F : X \to 2^Y$ is a closed convex multifunction, A is a closed convex subset of X, and b is a given point in Y. Recall that $S = \{x \in A :\ b \in F(x)\}$ is the solution set of the corresponding generalized equation with constraint (GEC).

Recently, in dealing with the inequality defined by a proper lower semicontinuous convex function, the authors [36] introduced and discussed the generalized BCQ and strong BCQ.

In terms of the coderivative replacing the subdifferential and the singular subdifferential, we can extend the concept of the generalized BCQ and strong BCQ to the case of a generalized equation with constraint (GEC). Explicitly, we say that (GEC) has the BCQ at $a \in S$ if

$$N(S, a) = D^*F(a, b)(Y^*) + N(A, a) \tag{3.1}$$

and (GEC) has the strong BCQ at $a \in S$ if there exists $\tau \in (0,\ +\infty)$ such that

$$N(S, a) \cap B_{X^*} \subset \tau(D^*F(a, b)(B_{Y^*}) + N(A, a)) \cap B_{X^*}. \tag{3.2}$$

The following theorem establishes the relationship between the metric subregularity and strong BCQ.

THEOREM 3.1. *Let $a \in S$. Then, the generalized equation (GEC) is metrically subregular at a if and only if there exist $\tau,\ \delta \in (0,\ +\infty)$ such that (GEC) has the strong BCQ at all points in $\mathrm{bd}(S) \cap B(a, \delta)$ with the same constant.*

Proof. Suppose that (GEC) is metrically subregular at a. Then there exist $\tau, \delta \in (0,\ +\infty)$ such that (1.2) holds. For any $(x, y) \in X \times Y$, let $\|(x, y)\|_\tau := \frac{\tau+1}{\tau}\|x\| + \|y\|$. Then $\|\cdot\|_\tau$ is a norm on $X \times Y$ inducing the product topology, and the unit ball of the dual space of $(X \times Y, \|\cdot\|_\tau)$ is $(\frac{\tau}{\tau+1}B_{X^*}) \times B_{Y^*}$. We claim that

$$d(x, S) \le \tau(d_{\|\cdot\|_\tau}((x, y), \mathrm{Gr}(F)) + \|y-b\| + d(x, A)) \quad \forall (x, y) \in B\left(a, \frac{\delta}{2}\right) \times Y, \tag{3.3}$$

where the distance $d_{\|\cdot\|_\tau}$ is with respect to the norm $\|\cdot\|_\tau$. Suppose to the contrary that (3.3) does not hold. Then there exists $(x_0, y_0) \in B(a, \frac{\delta}{2}) \times Y$ such that

$$d(x_0, S) > \tau[d_{\|\cdot\|_\tau}((x_0, y_0), \mathrm{Gr}(F)) + \|y_0 - b\| + d(x_0, A)].$$

It follows that there exists $u \in X$ such that

$$d(x_0, S) > \tau\left(\frac{\tau + 1}{\tau}\|u - x_0\| + d(y_0, Fu) + \|y_0 - b\| + d(x_0, A)\right),$$

and hence

$$d(x_0, S) > \|u - x_0\| + \tau(d(b, Fu) + d(u, A)).$$

Noting that

$$\|u - a\| \le \|u - x_0\| + \|x_0 - a\| < d(x_0, S) + \|x_0 - a\| \le 2\|x_0 - a\| < \delta,$$

it follows from (1.2) and the triangle inequality that

$$d(x_0, S) > \|u - x_0\| + d(u, S) \ge d(x_0, S),$$

which is a contradiction. Hence (3.3) holds.

We will establish the necessary part by showing that

$$N(S, z) \cap B_{X^*} \subset \tau(D^*F(z, b)(B_{Y^*}) + N(A, z) \cap B_{X^*}) \quad \forall z \in B\left(a, \frac{\delta}{2}\right) \cap S. \tag{3.4}$$

To do this, let $z \in S \cap B(a, \frac{\delta}{2})$ and $x^* \in N(S, z) \cap B_{X^*}$. Since A is convex, $N(S, z) \cap B_{X^*} = \partial d(\cdot, S)(z)$ (cf. [4, Theorem 1]). Thus,

$$\langle x^*, x - z\rangle \le d(x, S) - d(z, S) = d(x, S) \ \ \forall x \in X.$$

It follows from (3.3) that

$$\langle x^*, x - z\rangle \le \tau(d_{\|\cdot\|_\tau}((x, y), \mathrm{Gr}(F)) + \|y - b\| + d(x, A)) \quad \forall (x, y) \in B\left(z, \frac{\delta}{2} - \|z - a\|\right) \times Y.$$

Together with the convexity of F and A, this implies that $(\frac{x^*}{\tau}, 0) \in \partial\phi(z, b)$, where ϕ is the convex function defined by

$$\phi(x, y) := d_{\|\cdot\|_\tau}((x, y), \mathrm{Gr}(F)) + \|y - b\| + d(x, A) \quad \forall (x, y) \in X \times Y.$$

Noting that

$$\partial d_{\|\cdot\|_\tau}(\cdot, \mathrm{Gr}(F))(z, b) \subset N(\mathrm{Gr}(F), (z, b)),$$

it follows from [5, Proposition 2.3.2] that

$$\left(\frac{x^*}{\tau}, 0\right) \in N(\mathrm{Gr}(F), (z, b)) + \{0\} \times B_{Y^*} + (N(A, z) \cap B_{X^*}) \times \{0\}.$$

This implies that $x^* \in \tau(D^*F(z, b)(B_{Y^*}) + N(A, z) \cap B_{X^*})$, and hence that (3.4) holds, as is required to show.

Conversely, suppose that there exist $\tau', \delta' \in (0, +\infty)$ such that (GEC) has the strong BCQ at each point of $\mathrm{bd}(S) \cap B(a, \delta')$ with the constant τ'. Let $x \in B(a, \frac{\delta'}{2}) \setminus S$. Then, $d(x, S) \le \|x - a\| < \frac{\delta'}{2}$. Let $\beta \in (\frac{2d(x,S)}{\delta'}, 1)$. Then, by Lemma 2.1 there exists $u \in \mathrm{bd}(S)$ and $x^* \in N(S, u)$ with $\|x^*\| = 1$ such that $\beta\|x - u\| \le d(x, S)$ and

$$\beta\|x - u\| \le \langle x^*, x - u\rangle. \tag{3.5}$$

Thus, $\|x - u\| < \frac{\delta'}{2}$. Hence $\|u - a\| \le \|u - x\| + \|x - a\| < \delta'$, and so (GEC) has the strong BCQ at u with the constant τ. Therefore, there exists $y^* \in B_{Y^*}$,

$x_1^* \in D^*F(u,b)(y^*)$, and $x_2^* \in N(A,u) \cap B_{X^*} = \partial d(\cdot, A)(u)$ (by [4, Theorem 1]) such that $x^* = \tau'(x_1^* + x_2^*)$. By the convexity of F and A, one has

$$\langle x_1^*, x-u\rangle \le \langle y^*, y-b\rangle \quad \forall y \in F(x) \text{ and } \langle x_2^*, x-u\rangle \le d(x,A) - d(u,A) = d(x,A).$$

Hence,

$$\langle x^*, x-u\rangle \le \tau'(\langle y^*, y-b\rangle + d(x,A)) \le \tau'(\|y-b\| + d(x,A)) \quad \forall y \in F(x).$$

This and (3.5) imply that $\beta\|x-u\| \le \tau'(d(b,F(x)) + d(x,A))$. It follows from $u \in S$ that $\beta d(x,S) \le \tau'(d(b,F(x)) + d(x,A))$. Since β can be arbitrarily close to 1, $d(x,S) \le \tau'(d(b,F(x)) + d(x,A))$. This shows that (GEC) is metrically subregular at a. This completes the proof. □

Theorem 3.1 recaptures some earlier results dealing only with numerical valued functions. Let $f : X \to R \cup \{+\infty\}$ be a proper lower semicontinuous convex function. When $F(x) = [f(x), +\infty)$, $b = 0$, and $A = X$, then Theorem 3.1 is obtained in [36]. A slightly earlier result is due to Burke and Deng who showed in [3, Theorem 5.2] that if X is a Hilbert space, $F(x) = [f(x), +\infty)$, $b = \inf_{x\in X} f(x)$, and $A = X$, then (GEC) is metrically subregular at a if and only if there exists $\tau \in [0, +\infty)$ such that

$$N(S,x) \cap B_{X^*} \subset \tau \mathrm{cl}^*(\partial f(a)),$$

where cl^* denotes the weak* closure.

Remark 3.1. Let $\tau(F,a,b;A) := \inf\{\tau > 0 : \text{ (1.2) holds}\}$. For $u \in S$, let

$$\gamma(F,u,b;A) := \inf\{\tau > 0 : \text{ (GEC) has the strong BCQ at } u \text{ with the constant } \tau\}.$$

By the proof of Theorem 3.1, one can see that

$$\tau(F,a,b;A) = \limsup_{u \overset{\mathrm{bd}(S)}{\longrightarrow} a} \gamma(F,u,b;A) \ge \gamma(F,a,b;A). \tag{3.6}$$

In general, (GEC) is not necessarily metrically subregular at a if (GEC) has the strong BCQ only at a (see [36, Example 2]). But, when S is assumed to be "locally conical" at a, Theorem 3.1 and (3.6) can be sharpened. To do this, we need the following lemma.

LEMMA 3.1. *Let $s_1, s_2 \in S$ be such that $\langle u^*, s_1\rangle = \langle u^*, s_2\rangle$. Then,*

$$u^* \in D^*F(s_1,b)(B_{Y^*}) + N(A,s_1) \cap B_{X^*} \Leftrightarrow u^* \in D^*F(s_2,b)(B_{Y^*}) + N(A,s_2) \cap B_{X^*}.$$

Proof. Obviously we need only prove one direction of the implications, say " $\Rightarrow$." Let

$$\psi(x,y) := \|y-b\| + d(x,A) + \delta_{\mathrm{Gr}(F)}(x,y) \quad \forall (x,y) \in X \times Y,$$

where $\delta_{\mathrm{Gr}(F)}$ denotes the indicator function of $\mathrm{Gr}(F)$. It follows from [4, Theorem 1] and [5, Proposition 2.3.2] that

$$\partial\psi(s,b) = \{0\} \times B_{Y^*} + (N(A,s) \cap B_{X^*}) \times \{0\} + N(\mathrm{Gr}(F),(s,b)) \quad \forall s \in S. \tag{3.7}$$

Suppose that $u^* \in D^*F(s_1,b)(B_{Y^*}) + N(A,s_1) \cap B_{X^*}$. Then, by (3.7), one has $(u^*,0) \in \partial\psi(s_1,b)$. Hence,

$$\langle u^*, x-s_1\rangle \le \psi(x,y) - \psi(s_1,b) \quad \forall (x,y) \in X \times Y.$$

Since $\langle u^*, s_1\rangle = \langle u^*, s_2\rangle$ and $\psi(s_1, b) = \psi(s_2, b) = 0$,

$$\langle u^*, x - s_2\rangle \le \psi(x, y) - \psi(s_2, b) \quad \forall (x, y) \in X \times Y.$$

Therefore, $(u^*, 0) \in \partial\psi(s_2, b)$. It follows from (3.7) that $u^* \in D^*F(s_2, b)(B_{Y^*}) + N(A, s_2) \cap B_{X^*}$. This shows that the implication " $\Rightarrow$ " holds. Hence, the proof is completed. □

THEOREM 3.2. *Let $a \in S$. Suppose that there exist a cone C and a neighborhood V of a such that $S \cap V = (a + C) \cap V$. Then*

$$\tau(F, a, b; A) = \gamma(F, a, b; A).$$

Consequently, (GEC) is metrically subregular at a if and only if (GEC) has the strong BCQ at a.

Proof. In view of (3.6), we need only show that

$$\limsup_{u \overset{\mathrm{bd}(S)}{\longrightarrow} a} \gamma(F, u, b; A) \le \gamma(F, a, b; A). \tag{3.8}$$

Let $\delta > 0$ be such that $B(a, \delta) \subset V$. To prove (3.8), it suffices to show that for any $u \in S \cap B(a, \delta)$,

$$\gamma(F, u, b; A) \le \gamma(F, a, b; A). \tag{3.9}$$

We first show that

$$N(S, u) \subset N(S, a) \quad \forall u \in S \cap B(a, \delta). \tag{3.10}$$

To do this, let $u \in S \cap B(a, \delta)$ and $x^* \in N(S, u)$. Noting that V is a neighborhood of u, we have

$$N(S, u) = N(S \cap V, u) = N((a + C) \cap V, u) = N(a + C, u).$$

Choosing $c_u \in C$ such that $u = a + c_u$, it follows that

$$\langle x^*, a + c_u\rangle = \sup\{\langle x^*, a + c\rangle : \ c \in C\}.$$

Since C is a cone, it follows that $\langle x^*, c_u\rangle = 0$ and hence

$$\langle x^*, u\rangle = \langle x^*, a\rangle = \sup\{\langle x^*, a + c\rangle : \ c \in C\}. \tag{3.11}$$

This implies that $x^* \in N(a + C, a) = N(S, a)$. Therefore, (3.10) holds. Since (3.9) trivially holds if $\gamma(F, a, b; A) = +\infty$, we assume henceforth that $\gamma(F, a, b; A) < +\infty$. Let $r \in (\gamma(F, a, b; A), \ +\infty)$. Then,

$$N(S, a) \cap B_{X^*} \subset r(D^*F(a, b)(B_{Y^*}) + N(A, a) \cap B_{X^*}).$$

Let $u \in S \cap B(a, \delta)$ and $x^* \in N(S, u) \cap B_{X^*}$. By (3.10), one has

$$x^* \in r(D^*F(a, b)(B_{Y^*}) + N(A, a) \cap B_{X^*}).$$

It follows from (3.11) and Lemma 3.1 that $x^* \in r(D^*F(u, b)(B_{Y^*}) + N(A, u) \cap B_{X^*})$. Therefore,

$$N(S, u) \cap B_{X^*} \subset r(D^*F(u, b)(B_{Y^*}) + N(A, u) \cap B_{X^*}).$$

This implies that $\gamma(F,u,b;A)\le r$. Letting $r\to\gamma(F,a,b;A)$, one sees that (3.9) holds. Hence, the proof is completed. □

Remark 3.2. If the solution set S is a polyhedron, then for each $a\in S$ there exist a cone C and a neighborhood V of a such that $S\cap V=(a+C)\cap V$; in fact, in this case we can choose C to be the tangent cone of S at a.

THEOREM 3.3. *Let $a\in S$,*

$$\tau_1:=\inf\{\tau>0:\ d(x,a+T(S,a))\le\tau(d(b,F(x))+d(x,A))\ \ \forall x \text{ close to } a\},$$

and

$$\tau_2:=\inf\{\tau>0:\ d(h,T(S,a))\le\tau(d(0,DF(a,b)(h))+d(h,T(A,a)))\quad\forall h\in X\}.$$

Then

$$\tau_1=\tau_2=\gamma(F,a,b;A).$$

Moreover,

$$\tau_2<+\infty\Longrightarrow T(S,a)=T(A,a)\cap DF(a,b)^{-1}(0). \tag{3.12}$$

Consequently, (GEC) has the strong BCQ at a if and only if the sublinear generalized equation (with constraint)

$$0\in DF(a,b)(x)\ \text{ subject to } x\in T(A,a)$$

is metrically subregular at 0.

Proof. We first show that $\tau_1=\tau_2$. Let $h\in X$, $y\in DF(a,b)(h)$, $u\in T(A,a)$, and $\varepsilon>0$. Then, there exists $t>0$ small enough that

$$(h,y)\in\frac{\mathrm{Gr}(F)-(a,b)}{t}+\varepsilon B_X\times\varepsilon B_Y\ \text{ and }\ u\in\frac{A-a}{t}+\varepsilon B_X.$$

Therefore, there exists $z\in B_X$ such that

$$b+ty\in F(a+th+t\varepsilon z)+t\varepsilon B_Y\ \text{ and }\ a+tu\in A+t\varepsilon B_X.$$

This implies that

$$d(b,F(a+th+t\varepsilon z))\le t\|y\|+t\varepsilon\ \text{ and }\ d(a+th+t\varepsilon z,A)\le t\|h-u\|+2t\varepsilon.$$

Considering an arbitrary $\tau>\tau_1$ and noting that $t>0$ is small enough, it follows that

$$\begin{aligned}\tau t(\|y\|+\|h-u\|+3\varepsilon)&\ge d(a+th+t\varepsilon z,a+T(S,a))\\&\ge d(th,T(S,a))-t\varepsilon\\&=td(h,T(S,a))-t\varepsilon,\end{aligned}$$

where the last equality holds because $T(S,a)$ is a cone. Therefore,

$$d(h,T(S,a))\le\tau(d(0,DF(a,b)(h))+d(h,T(A,a)))+(3\tau+1)\varepsilon.$$

Letting $\varepsilon\to0$ and $\tau\to\tau_1$, one has

$$d(h,T(S,a))\le\tau_1(d(0,DF(a,b)(h))+d(h,T(A,a))).$$

Hence, $\tau_2 \le \tau_1$. Conversely, by the convexity of F, one has

$$\mathrm{Gr}(F) - (a,b) \subset T(\mathrm{Gr}(F),(a,b)) = \mathrm{Gr}(DF(a,b)).$$

Then, for any $x \in X$, $F(x) - b \subset DF(a,b)(x-a)$, and so

$$d(0, DF(a,b)(x-a)) \le d(b, F(x)).$$

On the other hand, the convexity of A implies that

$$d(x-a, T(A,a)) \le d(x-a, A-a) \le d(x,A).$$

Hence, for any $x \in X$,

$$d(x-a, T(S,a)) \le \tau_2(d(0, DF(a,b)(x-a)) + d(x-a, T(A,a))) \le \tau_2(d(b, F(x)) + d(x,A)).$$

Therefore $\tau_1 \le \tau_2$ and so $\tau_1 = \tau_2$ is shown. Next, we show that $\gamma(F,a,b;A) = \tau_2$. By the definition of τ_2, we have

$$d(x, T(S,a)) \le \tau_2(d(0, DF(a,b)(x)) + d(x, T(A,a))) \quad \forall x \in X.$$

In the case when $\tau_2 < +\infty$, this implies that $T(A,a) \cap DF(a,b)^{-1}(0) \subset T(S,a)$ and hence (3.12) is seen to hold, as the converse inclusion is easy to verify by the convexity of F and A. From (3.12) it is straightforward to verify that

$$\gamma(F,a,b;A) = \gamma(DF(a,b),0,0;T(A,a)) \text{ and } \tau_2 = \tau(DF(a,b),0,0;T(A,a)).$$

This and Theorem 3.2 imply that $\tau_2 = \gamma(F,a,b;A)$. In the case when $\tau_2 = +\infty$, suppose to the contrary that $\tau_2 \ne \gamma(F,a,b;A)$. Then, $\gamma(F,a,b;A) < +\infty$. Let $x \in X \setminus T(S,a)$ and $\beta \in (0,\ 1)$. By Lemma 2.1 there exist $u \in T(S,a)$ and $x^* \in N(T(S,a),u)$ such that

$$\|x^*\| = 1 \text{ and } \langle x^*, x-u\rangle \ge \beta\|x-u\|. \tag{3.13}$$

Noting that $N(T(S,a),u) \subset N(T(S,a),0)$ (because $T(S,a)$ is a closed convex cone), it follows that $x^* \in N(T(S,a),0) = N(S,a)$ and $\langle x^*, u\rangle = 0$. Take a fixed η in $(\gamma(F,a,b;A),\infty)$. Then there exist $y^* \in \eta B_{Y^*}$, $x_1^* \in D^*F(a,b)(y^*)$, and $x_2^* \in \eta N(A,a) \cap B_{X^*}$ such that $x^* = x_1^* + x_2^*$. Equipping the product space $X \times Y$ with norm $\|(x,y)\|_\eta = \frac{\eta}{1+\eta}\|x\| + \|y\|$ for all $(x,y) \in X \times Y$ and noting that the unit ball of the dual space of $(X \times Y, \|\cdot\|_\eta)$ is $(\frac{\eta+1}{\eta}B_{X^*}) \times B_{Y^*}$, it follows from (2.1) and the convexity of $DF(a,b)$ and A that

$$\begin{aligned}\frac{1}{\eta}(x_1^*, -y^*) &\in N(\mathrm{Gr}(F),(a,b)) \cap \left(\left(\frac{\eta+1}{\eta}B_{X^*}\right) \times B_{Y^*}\right)\\ &= N(\mathrm{Gr}(DF(a,b)),(0,0)) \cap \left(\left(\frac{\eta+1}{\eta}B_{X^*}\right) \times B_{Y^*}\right)\\ &= \partial d_{\|\cdot\|_\eta}(\cdot, \mathrm{Gr}(DF(a,b)))(0,0)\end{aligned}$$

and

$$\frac{1}{\eta}x_2^* \in N(A,a) \cap B_{X^*} = N(T(A,a),0) \cap B_{X^*} = \partial d(\cdot, T(A,a))(0).$$

Therefore,

$$\frac{1}{\eta}\langle x_1^*, x\rangle \le d_{\|\cdot\|_\eta}((x,0), \mathrm{Gr}(DF(a,b)) \le d(0, DF(a,b)(x))$$

and $\frac{1}{\eta}\langle x_2^*, x\rangle \le d(x, T(A,a))$. Noting that $\langle x^*, u\rangle = 0$, it follows from (3.13) that

$$\frac{\beta\|u-x\|}{\eta} \le d(0, DF(a,b)(x)) + d(x, T(A,a)).$$

Therefore, $\frac{\beta d(x,T(S,a))}{\eta} \le d(0, DF(a,b)(x)) + d(x, T(A,a))$. Letting $\beta \to 1$, one has

$$d(x, T(S,a)) \le \eta(d(0, DF(a,b)(x)) + d(x, T(A,a))).$$

This contradicts $\tau_2 = +\infty$. Hence, the proof is completed. □

Let $\phi : X \to R \cup \{+\infty\}$ be a proper lower semicontinuous convex function. Consider the special case when $A = X$ and $F(x) = [\phi(x), +\infty)$ for all $x \in X$. In this case, $N(A,x) = \{0\}$ for any $x \in A$. For $a \in \mathrm{dom}(\phi)$, let $\partial^\infty\phi(a)$ denote the singular subdifferential of ϕ at a, namely $\partial^\infty\phi(a) = D^*F(a, \phi(a))(0)$. It is easy to verify from the convexity of ϕ that $\mathrm{dom}(D^*F(a,\phi(a))) \subset R_+$. Thus, noting that

$$D^*F(a,\phi(a))(1) = \partial\phi(a) \text{ and } \partial\phi(a) = \partial\phi(a) + \partial^\infty\phi(a)$$

and adopting the convention that $R_+\partial\phi(a)$ and $[0,\ 1]\partial\phi(a)$ are $\{0\}$ if $\partial\phi(a) = \emptyset$, one has

$$\begin{aligned} D^*F(a,\phi(a))(R) &= D^*F(a,\phi(a))(0) \bigcup D^*F(a,\phi(a))(R_+ \setminus \{0\}) \\ &= D^*F(a,\phi(a))(0) \bigcup R_+ D^*F(a,\phi(a))(1) \\ &= \partial^\infty\phi(a) + R_+\partial\phi(a) \end{aligned}$$

and

$$D^*F(a,\phi(a))([-1,\ 1]) = \partial^\infty\phi(a) + [0,\ 1]\partial\phi(a).$$

Therefore, our definitions of the BCQ and strong BCQ for generalized equations are, respectively, natural generalizations of the BCQ and strong BCQ of a convex inequality system (cf. [18, 19, 20, 36]). Thus, Theorems 3.1 and 3.3 extend Theorems 2.2 and 2.3 in [36] from the setting of a convex inequality to that of a convex generalized equation with constraint.

Since the strong BCQ implies the BCQ, the following proposition shows that the converse also holds in some interesting cases.

PROPOSITION 3.4. *Let $a \in S$ and suppose that $N(S,a)$ is a polyhedron in a finite dimensional subspace of X^*. Then (GEC) has the BCQ at a if and only if it has the strong BCQ at a.*

Proof. We need only show the necessity part. Suppose that (GEC) has the BCQ at a. It suffices to show that there exists $\tau > 0$ such that

$$N(S,a) \cap B_{X^*} \subset \tau(D^*F(a,b)(B_{Y^*}) + N(A,a) \cap B_{X^*}). \tag{3.14}$$

Let E be a finite dimensional subspace of X^* such that $N(S,a) \subset E$. Let

$$L := N(S,a) \cap -N(S,a),$$

namely L is the largest subspace contained in $N(S,a)$. Take a subspace $L^\perp$ of E such that

$$L \cap L^\perp = \{0\} \text{ and } E = L + L^\perp. \tag{3.15}$$

Since $N(S,a)$ is a polyhedral cone in E, by [31, Theorem 19.1] there exists a polyhedron cone $C \subset L^\perp$ containing no lines such that

$$N(S,a) = C + L. \tag{3.16}$$

On the other hand, $\dim(E) < \infty$ and (3.15) imply that there exists $\delta \in (0, +\infty)$ such that $(C+L) \cap B_{X^*} \subset \delta(C \cap B_{X^*} + L \cap B_{X^*})$. It follows from (3.16) that

$$N(S,a) \cap B_{X^*} \subset \delta(C \cap B_{X^*} + L \cap B_{X^*}). \tag{3.17}$$

Since L is a finite dimensional space, there exist $l_1, \ldots, l_m \in L$ such that

$$B_{X*} \cap L \subset \text{co}(l_1, \ldots, l_m). \tag{3.18}$$

Take $c_1, \ldots, c_n \in C$ such that $C = R^+\text{co}(c_1, \ldots, c_n)$ and $0 \notin \text{co}(c_1, \ldots, c_n)$ (because C is a finite dimensional polyhedron cone containing no lines). Without loss of generality, we assume that $\text{co}(c_1, \ldots, c_n) \cap B_{X^*} = \emptyset$. We note that

$$C \cap B_{X*} \subset \text{co}(0, c_1, \ldots, c_n). \tag{3.19}$$

By (3.16) and the BCQ assumption, there exist

$$\{y_1^*, \ldots, y_n^*, \tilde{y}_1^*, \ldots, \tilde{y}_m^*\} \subset Y^* \text{ and } \{a_1^*, \ldots, a_n^*, \tilde{a}_1^*, \ldots, \tilde{a}_m^*\} \subset N(A,a)$$

such that

$$c_i \in D^*F(a,b)(y_i^*) + a_i^*,\ 1 \le i \le n \text{ and } l_j \in D^*F(a,b)(\tilde{y}_j^*) + \tilde{a}_j^*,\ 1 \le j \le m.$$

Let $\kappa := \max_{1 \le i \le n, 1 \le j \le m} \|y_i^*\| + \|a_i^*\| + \|\tilde{y}_j^*\| + \|\tilde{a}_j^*\|$. It follows from (3.18) and (3.19) that

$$C \cap B_{X^*} + L \cap B_{X^*} \subset \kappa(D^*F(a,b)(B_{Y^*}) + N(A,a) \cap B_{X^*}).$$

This and (3.17) imply that (3.14) holds with $\tau = \delta\kappa$. □

COROLLARY 3.5. *Let $f_1, \ldots, f_n : X \to R \cup \{+\infty\}$ be proper lower semicontinuous convex functions and consider generalized equation (GEC) with $A = X$, $Y = R^n$, $b = (b_1, \ldots, b_n) \in R^n$, and F being defined by*

$$F(x) = (f_1(x), \ldots, f_n(x)) + R_+^n \quad \forall x \in X.$$

Suppose that each f_i is differentiable at $a \in S$. Then, for the said generalized equation, the BCQ and strong BCQ are equivalent at a.

Proof. In view of Proposition 3.4, it suffices to show that

$$D^*F(a,b)(R^n) = R_+\text{co}\{f_i'(a) :\ i \in J(a)\}, \tag{3.20}$$

where $J(a) := \{1 \le i \le n :\ f_i(a) = b_i\}$. To do this, we first note that $\text{dom}(D^*F(a,b)) = R_+^n$ (because each convex function f_i is differentiable at a). Let $(r_1, \ldots, r_n) \in R_+^n \setminus \{0\}$ and $x^* \in D^*F(a,b)(r_1, \ldots, r_n)$. Then

$$\langle x^*, x\rangle - \sum_{i=1}^n r_i(f_i(x) + t_i) \le \langle x^*, a\rangle - \sum_{i=1}^n r_i b_i$$

for any $x \in X$ and $(t_1, \ldots, t_n) \in R^n_+$. Noting that $f_i(a) = b_i$ for any $i \in J(a)$, $f_i(a) < b_i$ for any $i \notin J(a)$, and $a \in \text{int}(\text{dom}(f_i))$ for $1 \le i \le n$, it follows that $r_i = 0$ for any $i \notin J(a)$ and

$$\langle x^*, x - a\rangle \le \sum_{i\in J(a)} r_i f(x) - \sum_{i\in J(a)} r_i f_i(a)$$

for all $x \in X$. This implies that $x^* = \sum_{i\in J(a)} r_i f_i'(a)$. Thus, $x^* \in R_+\text{co}\{f_i'(a) : i \in J(a)\}$. This shows that $D^*F(a,b)(R^n) \subset R_+\text{co}\{f_i'(a) : \ i \in J(a)\}$. Conversely, let $x^* \in R_+\text{co}\{f_i'(a) : \ i \in J(a)\}$. Then there exists $(c_1, \ldots, c_n) \in R^n_+$ with $c_i = 0$ for all $i \notin J(a)$ such that $x^* = \sum_{i=1}^n c_i f_i'(a)$. Noting that for each i,

$$\langle c_i f_i'(a), x - a\rangle \le c_i(f_i(x) + t_i - b_i) \quad \forall x \in X \text{ and } \forall t_i \ge 0,$$

it follows that $x^* \in D^*F(a,b)(c_1, \ldots, c_n)$. This shows that

$$R_+\text{co}\{f_i'(a) : \ i \in J(a)\} \subset D^*F(a,b)(R^n).$$

Hence, (3.20) holds. □

4. Calmness of convex multifunctions. Throughout this section, let $M : Y \to 2^X$ be a closed convex multifunction and A be a closed convex subset of X. Let $\bar{y} \in Y$ and $\bar{x} \in M(\bar{y}) \cap A$.

Recall (cf. [8, 9, 10] and [15]) that M is said to be calm at $(\bar{y}, \bar{x})$ if there exists a constant Ł > 0 such that

$$d(x, M(\bar{y})) \le L\|y - \bar{y}\| \quad \forall (y,x) \in \text{Gr}(M) \text{ close to } (\bar{y}, \bar{x}). \tag{4.1}$$

More generally, M is said to be calm at $(\bar{y}, \bar{x})$ over A if there exists a constant Ł > 0 such that

$$d(x, M(\bar{y}) \cap A) \le L(\|y - \bar{y}\| + d(x, A)) \quad \forall (y,x) \in \text{Gr}(M) \text{ close to } (\bar{y}, \bar{x}). \tag{4.2}$$

Let $\tilde{M} : Y \times X \to 2^X$ be defined by

$$\tilde{M}(y,z) = M(y) \cap (-z + A) \text{ for any } (y,z) \in Y \times X$$

and $Y \times X$ be equipped with the norm $\|(y,z)\| = \|y\| + \|z\|$ for any $(y,z) \in Y \times X$. Then, as observed by one of the referees, (4.2) holds if and only if

$$d(x, \tilde{M}(\bar{y}, 0)) \le L\|(y,z) - (\bar{y}, 0)\| \quad \forall (y,z;x) \in \text{Gr}(\tilde{M}) \text{ close to } (\bar{y}, 0; \bar{x}).$$

Hence, M is calm at $(\bar{y}, \bar{x})$ over A if and only if $\tilde{M}$ is calm at $(\bar{y}, 0; \bar{x})$. A more general intersection map has been studied by Klatte and Kummer [16].

Since $d(x, \emptyset) = +\infty$ and $d(x, M(\bar{y})) \le \|x - \bar{x}\|$, it is easy to verify that (4.2) holds if and only if

$$d(x, M(\bar{y}) \cap A) \le L(d(\bar{y}, M^{-1}(x)) + d(x, A)) \quad \forall x \text{ close to } \bar{x}. \tag{4.3}$$

Letting $b = \bar{y}$ and $F(x) = M^{-1}(x)$, it follows that (GEC) is metrically subregular at $\bar{x}$ if and only if M is calm at $(\bar{y}, \bar{x})$ over A. Thus, by Theorems 3.1 and 3.3, we have the following results.

THEOREM 4.1. *The following statements are equivalent.*

(i) *M is calm at $(\bar{y}, \bar{x})$ over A.*
(ii) *There exist $\tau, \delta \in (0, +\infty)$ such that for all $u \in B(\bar{x}, \delta) \cap \mathrm{bd}(M(\bar{y}) \cap A)$,*

$$N(M(\bar{y}) \cap A, u) \cap B_{X^*} \subset \tau(D^*M^{-1}(\bar{x}, \bar{y})(B_{Y^*}) + N(A, u) \cap B_{X^*}).$$

(iii) *There exists $\delta \in (0, +\infty)$ such that for all $u \in \mathrm{bd}(M(\bar{y}) \cap A)$ close to $\bar{x}$, the tangent derivative $DM(\bar{y}, u)$ is calm at $(0, 0)$ over $T(A, u)$ with the same constant.*

Using Theorem 4.1, we can establish some characterization of the strong calmness: M is called strongly calm at $(\bar{y}, \bar{x})$ over A if there exists $L \in [0, +\infty)$ such that

$$\|x - \bar{x}\| \le L(\|y - \bar{y}\| + d(x, A)) \quad \forall (y, x) \in \mathrm{Gr}(M) \text{ close to } (\bar{y}, \bar{x}). \tag{4.4}$$

When $A = X$, the strong calmness means the local upper Lipschitz property presented in [15, p. 6]. From the convexity of $M(\bar{y}) \cap A$, it is clear that M is strongly calm at $(\bar{y}, \bar{x})$ over A if and only if $M(\bar{y}) \cap A = \{\bar{x}\}$ and M is calm at $(\bar{y}, \bar{x})$ over A.

Corollary 4.2. *The following statements are equivalent.*
(i) *M is strongly calm at $(\bar{y}, \bar{x})$ over A.*
(ii) *There exists $L \in [0, +\infty)$ such that*

$$\|x - \bar{x}\| \le L(\|y - \bar{y}\| + d(x, A)) \quad \forall (y, x) \in \mathrm{Gr}(M).$$

(iii) *The tangent derivative $DM(\bar{y}, \bar{x})$ is strongly calm at $(0, 0)$ over $T(A, \bar{x})$.*
(iv) $0 \in \mathrm{int}(D^*M^{-1}(\bar{x}, \bar{y})(Y^*) + N(A, \bar{x}))$.
(v) *There exists $r > 0$ such that*

$$rB_{X^*} \subset D^*M^{-1}(\bar{x}, \bar{y})(B_{Y^*}) + N(A, \bar{x}) \cap B_{X^*}.$$

Proof. First, we show that (i)⇔(v). By the evident fact $N(\{\bar{x}\}, \bar{x}) = X^*$ and by Theorem 4.1, we need only show that (v)⇒$M(\bar{y}) \cap A = \{\bar{x}\}$. Take an arbitrary $x \in M(\bar{y}) \cap A$ and $x^* \in B_{X^*}$ such that $\|x - \bar{x}\| = \langle x^*, x - \bar{x}\rangle$. By (v), there exist $y^* \in B_{Y^*}$, $x_1^* \in D^*M^{-1}(\bar{x}, \bar{y})(y^*)$ and $x_2^* \in N(A, \bar{x}) \cap B_{X^*}$ such that $rx^* = x_1^* + x_2^*$. Hence,

$$r\|x - \bar{x}\| = \langle x_1^*, x - \bar{x}\rangle + \langle x_2^*, x - \bar{x}\rangle.$$

Noting that $\langle x_1^*, x - \bar{x}\rangle \le \langle y^*, \bar{y} - \bar{y}\rangle = 0$ and $\langle x_2^*, x - \bar{x}\rangle \le 0$, it follows that $r\|x - \bar{x}\| \le 0$ for any $x \in M(\bar{y}) \cap A$. This shows that $M(\bar{y}) \cap A = \{\bar{x}\}$.

Noting that $D^*M^{-1}(\bar{x}, \bar{y}) = D^*(DM(\bar{y}, \bar{x}))^{-1}(0, 0)$, (iii)⇔(v) is immediate from (i)⇔(v).

It is clear that (ii)⟹(i) and (v)⟹(iv).

Suppose that (i) holds. Then there exists $L \in [0, +\infty)$ such that (4.4) holds. Let (y, x) be an arbitrary element in $\mathrm{Gr}(M)$ and $t \in (0, 1)$ be small enough such that $(ty + (1 - t)\bar{y}, tx + (1 - t)\bar{x})$ close enough to $(\bar{y}, \bar{x})$. By (4.4) and the convexity of M, one has

$$\|tx + (1 - t)\bar{x} - \bar{x}\| \le L(\|ty + (1 - t)\bar{y} - \bar{y}\| + d(tx + (1 - t)\bar{x}, A)).$$

It follows from the convexity of A and $\bar{x} \in A$ that $\|x - \bar{x}\| \le L(\|y - \bar{y}\| + d(x, A))$. This shows that (i)⟹(ii).

It remains to show that (iv)⇒(v). Suppose that (iv) holds. Since $N(A, x)$ and $D^*M^{-1}(x, \bar{y})(Y^*)$ are cones,

$$X^* = D^*M^{-1}(\bar{x}, \bar{y})(Y^*) + N(A, \bar{x}) = \bigcup_{n=1}^{\infty} (D^*M^{-1}(\bar{x}, \bar{y})(nB_{Y^*}) + N(A, \bar{x}) \cap nB_{X^*}).$$

Noting that, by the Alaoglu theorem, each set $D^*M^{-1}(\bar{x},\bar{y})(nB_{Y^*})+N(A,\bar{x})\cap nB_{X^*}$ is weak* closed, it follows from the well known Baire category theorem and (iv) that

$$0\in \operatorname{int}(D^*M^{-1}(\bar{x},\bar{y})(B_{Y^*})+N(A,\bar{x})\cap B_{X^*}).$$

Hence there exists $r>0$ such that (v) holds. Hence, the proof is completed. □

Remark. In a recent paper [8], Henrion and Jourani considered the calmness of the convex multifunction M_0 of the following type:

$$M_0(y)=\{x\in C:\ f(x)\le y\}\quad \forall y\in R, \tag{4.5}$$

where C is a closed convex subset of X and $f: X\to R\cup\{+\infty\}$ is a proper lower semicontinuous convex function. In particular, as a main result, they established the following result.

THEOREM HJ (see [8, Theorem 3.3]). Let M_0 be defined by (4.5). Then M_0 is calm at $(0,\bar{x})\in \operatorname{Gr}(M_0)$ if one of the following conditions is satisfied:

(C1) $f(\bar{x})<0$,

(C2) $\operatorname{bd}\partial f(\bar{x})\cap \operatorname{bd}N(C,\bar{x})\ne \partial f(\bar{x})\cap N(C,\bar{x})$,

(C3) $\operatorname{bd}\partial f(\bar{x})\cap \operatorname{bd}N(C,\bar{x})=\emptyset$ and (CD*) (see [8] for the definition of condition (CD*)).

As observed by Henrion and Jourani [8], (C3)$\Longrightarrow$either (C2) or (C1) and

$$\text{(C2)}\Longrightarrow \operatorname{int}\partial f(\bar{x})\cap -N(C,\bar{x})\ne\emptyset \text{ or } \partial f(\bar{x})\cap -\operatorname{int}N(C,\bar{x})\ne\emptyset.$$

Considering that (C1)$\Longrightarrow$the calmness of M_0 at $(0,\bar{x})$ is an immediate consequence of the Robinson–Ursescu theorem (cf. [29, 33]), the main part of Theorem HJ can be rewritten as follows: M_0 is calm at $(0,\bar{x})\in \operatorname{Gr}(M_0)$ if

$$\operatorname{int}(\partial f(\bar{x}))\cap -N(C,\bar{x})\ne\emptyset \ \text{ or } \ \partial f(\bar{x})\cap -\operatorname{int}N(C,\bar{x})\ne\emptyset. \tag{4.6}$$

Let $A=C$, $Y=R$ and $M(y)=\{x\in X:\ f(x)\le y\}$ for all $y\in Y$. It is clear that M_0 is calm at $(0,\bar{x})$ if M is calm at $(0,\bar{x})$ over A. Since

$$\partial f(\bar{x})=D^*M^{-1}(\bar{x},f(\bar{x}))(1)\subset D^*M^{-1}(\bar{x},f(\bar{x}))(Y^*),$$

we see immediately that (iv) in Corollary 4.2 holds whenever (4.6) holds. On the other hand, below is a simple example for which Corollary 4.2 is applicable but Theorem HJ is not.

Let $X=R^2$, $C=R\times\{0\}$, $\bar{x}=(0,0)$, and $f(x_1,x_2)=|x_1|$ for all $(x_1,x_2)\in R^2$. Then (4.6) is not satisfied because

$$M_0(0)=\{\bar{x}\},\ \partial f(\bar{x})=[-1,\ 1]\times\{0\} \text{ and } N(C,\bar{x})=\{0\}\times R.$$

However, Corollary 4.2 is applicable because (iv) holds as $D^*M^{-1}(\bar{x},f(\bar{x}))(Y^*)=R\times\{0\}$ and so $D^*M^{-1}(\bar{x},f(\bar{x}))(Y^*)+N(C,\bar{x})=R^2$.

The calmness modulus of M at $(\bar{y},\bar{x})$ is denoted by $\eta(M;\bar{y},\bar{x})$ and is defined by

$$\eta(M;\bar{y},\bar{x}):=\inf\{L\in(0,\ +\infty):\ (4.1)\text{ holds}\}.$$

As applications of Theorems 3.1 and 3.2, we establish formulas representing $\eta(M;\bar{y},\bar{x})$.

THEOREM 4.3. $\eta(M;\bar{y},\bar{x})=\limsup_{u\overset{\operatorname{bd}(M(\bar{y}))}{\longrightarrow}\bar{x}}\|D^*M(\bar{y},u)|_{-N(M(\bar{y}),u)}\|^-$.

Proof. Let $F = M^{-1}$ and $A = X$. Since (4.2)$\Leftrightarrow$(4.3), $\eta(M; \bar{y}, \bar{x}) = \tau(F, \bar{x}, \bar{y}; A)$. By (3.6), it suffices to show that

$$\gamma(F, u, \bar{y}; A) = \|D^*M(\bar{y}, u)|_{-N(M(\bar{y}),u)}\|^- \quad \forall u \in F^{-1}(\bar{y}) = M(\bar{y}). \tag{4.7}$$

Let $u \in F^{-1}(\bar{y}) = M(\bar{y})$. By definitions, it is clear that

$$x^* \in D^*F(u, \bar{y})(y^*) \Longleftrightarrow -y^* \in D^*M(\bar{y}, u)(-x^*).$$

Let $\tau > \gamma(F, u, \bar{y}; A)$. Noting that $N(A, u) = N(X, u) = \{0\}$, one has

$$N(M(\bar{y}), u) \cap B_{X^*} \subset D^*F(u, \bar{y})(\tau B_{Y^*}).$$

Hence, for any $x^* \in N(M(\bar{y}), u) \cap B_{X^*}$ there exists $y^* \in B_{Y^*}$ such that $x^* \in D^*F(u, \bar{y})(\tau y^*)$, that is, $-\tau y^* \in D^*M(\bar{y}, u)(-x^*)$. It follows that

$$\|D^*M(\bar{y}, u)|_{-N(M(\bar{y}),u)}\|^- \le \tau.$$

Therefore, $\|D^*M(\bar{y}, u)|_{-N(M(\bar{y}),u)}\|^- \le \gamma(F, u, \bar{y}; A)$. To prove the converse inequality, let $\tau > \|D^*M(\bar{y}, u)|_{-N(M(\bar{y}),u)}\|^-$ and take $x^* \in N(F^{-1}(\bar{y}), u) = N(M(\bar{y}), u)$. Then, there exists $y^* \in D^*M(\bar{y}, u)(-x^*)$ such that $\|y^*\| < \tau$; this means $x^* \in D^*F(u, \bar{y})(\tau B_{Y^*})$. Hence, $N(F^{-1}(\bar{y}), u) \cap B_{X^*} \subset D^*F(u, \bar{y})(\tau B_{Y^*})$. It follows that $\gamma(F, u, \bar{y}; A) \le \tau$. Therefore, $\gamma(F, u, \bar{y}; A) \le \|D^*M(\bar{y}, u)|_{-N(M(\bar{y}),u)}\|^-$. This shows that (4.7) holds. □

Remark. In contrast to formula (1.4) of the modulus of the metric regularity, $\eta(M; \bar{y}, \bar{x})$ is not necessarily equivalent to $\|D^*M(\bar{y}, \bar{x})|_{-N(M(\bar{y}),\bar{x})}\|^-$ even when $Y = R^2$ and $X = R$ (cf. [36, p. 763, Example 2]). Nevertheless, the following theorem shows an interesting case for which the equality holds.

THEOREM 4.4. *Suppose that there exist a cone C and a neighborhood V of $\bar{x}$ such that $M(\bar{y}) \cap V = (\bar{x} + C) \cap V$. Then $\eta(M; \bar{y}, \bar{x}) = \|D^*M(\bar{y}, \bar{x})|_{-N(M(\bar{y}),\bar{x})}\|^-$.*

The proof of Theorem 4.4 is similar to that of Theorem 4.3 but using Theorem 3.2 in place of Theorem 3.1.

5. Recession core and global metric subregularity. Let K be a closed convex subset of X.

Recall that $e \in K$ is called an extreme point of K if $x_1 = x_2$ whenever $e = tx_1 + (1 - t)x_2$ with $x_1, x_2 \in K$ and $t \in (0,\ 1)$. We denote by $\mathrm{ext}(K)$ the set of all extreme points of K (usually $\mathrm{ext}(K)$ is called the extreme boundary of K).

Let K^∞ denote the recession cone of K, that is,

$$K^\infty := \{h \in X :\ K + th \subset K \quad \forall t \ge 0\}.$$

It is known that K^∞ is a closed convex cone, and

$$\begin{aligned} K^\infty &= \{h \in X :\ x + R_+h \subset K \text{ for some } x \in K\} \\ &= \{h \in X :\ \exists x_n \in K \text{ and } \exists t_n > 0 \text{ such that } t_n \to 0 \text{ and } t_n x_n \to h\}. \end{aligned}$$

Clearly, $K + K^\infty = K$. It is well known that if K is a closed convex subset of R^n containing no lines, then $K = \mathrm{co}(\mathrm{ext}(K)) + K^\infty$.

As a generalization of $\mathrm{co}(\mathrm{ext}(K))$, the authors [28] introduced the concept of recession property: a convex subset A of K is said to have the recession property if $K = A + K^\infty$.

Simplifying the recession property, we can now give a generalization of ext(K): a subset C of K is said to be a recession core of K if

$$K = \mathrm{co}(C) + K^\infty. \tag{5.1}$$

Thus, C is a recession core of K if and only if co(C) is a subset of K with recession property.

Let "$\leq_{K^\infty}$" denote the order induced by the cone K^∞, that is, $x_1 \leq_{K^\infty} x_2$ if and only if $x_2 - x_1 \in K^\infty$.

Let A be a subset of X. We say that $a \in A$ is a minimal element of A with respect to "$\leq_{K^\infty}$" if $a \leq_{K^\infty} x$ whenever $x \in A$ and $x \leq_{K^\infty} a$. We denote by $\mathrm{Min}(A, K^\infty)$ the set of all minimal elements of A.

Let

$$\mathrm{Lin}(K) := K^\infty \cap -K^\infty.$$

Then

$$\mathrm{Lin}(K) = \{h \in X : \ K + Rh \subset K\} = \{h \in X : \ x + Rh \subset K \text{ for some } x \in K\}.$$

Moreover, $K = K + \mathrm{Lin}(K)$ and $L \subset \mathrm{Lin}(K)$ whenever L is a subspace of X such that $K = K + L$. Therefore, K contains no lines if and only if $\mathrm{Lin}(K) = \{0\}$.

PROPOSITION 5.1. *Let X be a reflexive Banach space and K a closed convex nonempty subset of X. Suppose that there exists a closed convex bounded set Θ such that*

$$\mathrm{Lin}(K) \cap \Theta = \emptyset \ \text{ and } \ K^\infty = \mathrm{Lin}(K) + R_+\Theta. \tag{5.2}$$

Then

$$K = \mathrm{Min}(K, K^\infty) + K^\infty. \tag{5.3}$$

In particular, $\mathrm{Min}(K, K^\infty)$ *is a recession core of K.*

Proof. By the reflexivity of X, the bounded closed convex set Θ is weakly compact. Letting $K_0 := R_+\Theta$, it follows that K_0 is a closed convex pointed cone. Let x be an arbitrary point in K. We claim that $K \cap (x - K_0)$ is bounded. If this is not the case, then there exist a sequence $\{s_n\}$ in R_+ and a sequence $\{\theta_n\}$ in Θ such that $s_n \to \infty$ and $x - s_n\theta_n \in K$ for all n. Thus, for any $t > 0$,

$$x - t\theta_n = \left(1 - \frac{t}{s_n}\right) x + \frac{t}{s_n}(x - s_n\theta_n) \in K \quad \forall n \text{ large enough.}$$

By the weak compactness of Θ, without loss of generality we can assume that $\{\theta_n\}$ converges weakly to some $\theta \in \Theta$. Therefore, $x - t\theta \in K$ for any $t \geq 0$. This implies that $-\theta \in K^\infty$. On the other hand, by the second equality in (5.2), one has $\theta \in K^\infty$ and so $\theta \in \mathrm{Lin}(K)$. This contradicts the first equality in (5.2) and therefore $K \cap (x - K_0)$ must be bounded (and hence weakly compact). It follows from [14, Corollary 3.1.16]) that $\mathrm{Min}(K \cap (x - K_0), K_0) \neq \emptyset$. Take $x' \in \mathrm{Min}(K \cap (x - K_0), K_0)$. Then $x' \in \mathrm{Min}(K, K_0)$ and $x \in x' + K_0$. Hence $K \subset \mathrm{Min}(K, K_0) + K_0$. Now to show (5.3), it suffices to show that $\mathrm{Min}(K, K_0) \subset \mathrm{Min}(K, K^\infty)$. Let $z \in \mathrm{Min}(K, K_0)$ and $y \in K$ with $y \leq_{K^\infty} z$. Then $z - y \in K^\infty$. By the second equality of (5.2), there exists $e \in \mathrm{Lin}(K)$ such that $z - y - e \in K_0$, that is, $y + e \leq_{K_0} z$. Noting that $y + e \in K$,

one has that $y + e - z \in K_0$ and hence $y - z \in \mathrm{Lin}(K) + K_0 = K^\infty$. This shows that $z \in \mathrm{Min}(K, K^\infty)$. Hence, the proof is completed. □

In the case when X is finite dimensional, the assumption made in (5.2) automatically holds (by $K = \mathrm{Lin}(K) + C$ and Klee's Theorem (cf. [14]), where C is a closed convex pointed cone). The following example shows that the reflexivity of X cannot be removed in Proposition 5.1.

Example. Let $X = l^1$ and $K = \{x = (t_1, t_2, \cdots) \in l^1 : \ t_n \geq -n \text{ for all } n\}$. It is easy to verify that $\mathrm{Lin}(K) = \{0\}$ and $K^\infty = \{x = (t_1, t_2, \cdots) \in l^1 : \ t_n \geq 0 \text{ for all } n\}$. Let $\Theta := \{x = (t_1, t_2, \cdots) \in K^\infty : \ \sum_{n=1}^\infty t_n = 1\}$. Then, Θ is a bounded closed convex set, $\mathrm{Lin}(K) \cap \Theta = \emptyset$ and $K^\infty = R^+\Theta$. But $\mathrm{Min}(K, K^\infty) = \emptyset$, and hence $K \neq \mathrm{Min}(K, K^\infty) + K^\infty$. Indeed, let $x = (t_1, t_2, \cdots)$ be any point in K. Noting that $\sum_{n=1}^\infty |t_n| < \infty$, there exists a natural number n_0 such that $|t_n| < n_0$ for all $n \geq n_0$. Take $x_0 = (s_1, s_2, \cdots)$ to satisfy $s_n = t_n$ for any $n \neq n_0$ and $s_{n_0} = -n_0$. It is clear that $x_0 \in K \setminus \{x\}$ and $x - x_0 \in K^\infty$. It follows that $x \notin \mathrm{Min}(K, K^\infty)$. This shows that $\mathrm{Min}(K, K^\infty) = \emptyset$.

It is clear that K has no extreme points if K contains lines. This motivates us to introduce a new concept—what we shall refer to as "generalized extreme points." Let X be a Hilbert space. For a closed convex subset K of X, let

$$\mathrm{Lin}(K)^\perp := \{x \in X : \ \langle x, y\rangle = 0 \quad \forall y \in \mathrm{Lin}(K)\}.$$

We say that e is a generalized extreme point of K if $e \in K \cap \mathrm{Lin}(K)^\perp$ and

$$x_1, x_2 \in K \text{ and } e = \frac{x_1 + x_2}{2} \Longrightarrow x_1 - x_2 \in \mathrm{Lin}(K). \tag{5.4}$$

We denote by $\mathrm{ext}_E(K)$ the set of all generalized extreme points of K. Clearly, $\mathrm{ext}_E(K) = \mathrm{ext}(K)$ if K contains no lines (i.e., $\mathrm{Lin}(K) = \{0\}$). Moreover, one has that

$$\mathrm{ext}_E(K) \subset \mathrm{Min}(K, K^\infty). \tag{5.5}$$

To see this, let $e \in \mathrm{ext}_E(K)$ and $x \in K$ with $x \leq_{K^\infty} e$. Then $e - x \in K^\infty$ and hence $2e - x = e + (e - x) \in K$. Since $e = \frac{x + (2e - x)}{2}$, it follows that $2(e - x) \in \mathrm{Lin}(K)$, which means $x - e \in \mathrm{Lin}(K) \subset K^\infty$. Hence $e \leq_{K^\infty} x$. This shows that $e \in \mathrm{Min}(K, K^\infty)$. Thus (5.5) is true.

PROPOSITION 5.2. *Let X be a Hilbert space and K a closed convex subset of X. Then*

$$\mathrm{ext}_E(K) = \mathrm{ext}(K \cap \mathrm{Lin}(K)^\perp). \tag{5.6}$$

Proof. By definition it is clear that $\mathrm{ext}_E(K) \subset \mathrm{ext}(K \cap \mathrm{Lin}(K)^\perp)$. Conversely, let $e \in \mathrm{ext}(K \cap \mathrm{Lin}(K)^\perp)$ and $x_1, x_2 \in K$ satisfy $e = \frac{x_1 + x_2}{2}$. Noting that for each $x \in X$ there exists a unique pair $(u, v) \in \mathrm{Lin}(K) \times \mathrm{Lin}(K)^\perp$ such that $x = u + v$, take $(u_1, v_1), (u_2, v_2) \in \mathrm{Lin}(K) \times \mathrm{Lin}(K)^\perp$ such that $x_1 = u_1 + v_1$ and $x_2 = u_2 + v_2$. Then $e = \frac{u_1 + u_2}{2} + \frac{v_1 + v_2}{2}$. It follows from $e \in \mathrm{ext}(K \cap \mathrm{Lin}(K)^\perp)$ that $u_1 + u_2 = 0$ and $e = v_1 = v_2$. Thus, $x_1 - x_2 = u_1 - u_2 \in \mathrm{Lin}(K)$ and hence $e \in \mathrm{ext}_E(K)$. This shows that $\mathrm{ext}_E(K) \supset \mathrm{ext}(K \cap \mathrm{Lin}(K)^\perp)$ and (5.6) is proved. □

PROPOSITION 5.3. *Let K be a closed convex subset of a Hilbert space X and $\prod$ be the project operator to $\mathrm{Lin}(K)^\perp$. Suppose that C is a recession core of K. Then $\mathrm{ext}_E(K) \subset \prod(C)$.*

Proof. Let $e \in \text{ext}_E(K)$. Then, by (5.1) there exist $x_1, \ldots, x_n \in C$, $t_1, \ldots, t_n \in [0,\ 1]$ with $\sum_{i=1}^n t_i = 1$, and $h \in K^\infty$ such that $e = \sum_{i=1}^n t_i x_i + h$. Then,

$$e = \frac{\left(\sum_{i=1}^n t_i x_i + \frac{1}{2}h\right) + \left(\sum_{i=1}^n t_i x_i + \frac{3}{2}h\right)}{2}.$$

Thus, by (5.4), one has $h \in \text{Lin}(K)$. It follows from the linearity of $\prod$ that

$$e = \prod(e) = \prod\left(\sum_{i=1}^n t_i x_i + h\right) = \sum_{i=1}^n t_i \prod(x_i) \in \prod(C).$$

This shows that $\text{ext}_E(K) \subset \prod(C)$. □

The following proposition shows that $\text{ext}_E(K)$ is a recession core of K when X is finite dimensional.

PROPOSITION 5.4. *Let K be a closed convex nonempty subset of R^n. Then*

$$K = \text{co}(\text{ext}_E(K)) + K^\infty. \tag{5.7}$$

Proof. Let $h \in R^n$ be such that $x + Rh \subset K \cap \text{Lin}(K)^\perp$ for some $x \in R^n$. Then $h \in \text{Lin}(K)$, and hence $\langle x + th, h\rangle = 0$ for all $t \in R$. It follows that $h = 0$. Therefore $K \cap \text{Lin}(K)^\perp$ is a closed convex subset containing no lines. It follows from [31, Theorem 18.5] that

$$K \cap \text{Lin}(K)^\perp = \text{co}(\text{ext}(K \cap \text{Lin}(K)^\perp)) + (K \cap \text{Lin}(K)^\perp)^\infty.$$

This and Proposition 5.2 imply that $K \cap \text{Lin}(K)^\perp \subset \text{co}(\text{ext}_E(K)) + K^\infty$. Noting that $K^\infty + \text{Lin}(K) = K^\infty$, one then has

$$K \cap \text{Lin}(K)^\perp + \text{Lin}(K) \subset \text{co}(\text{ext}_E(K)) + K^\infty. \tag{5.8}$$

Let $x \in K$ and take $x_1 \in \text{Lin}(K)$ and $x_2 \in \text{Lin}(K)^\perp$ such that $x = x_1 + x_2$. Hence $x_2 \in x + \text{Lin}(K) \subset K$ and so $x_2 \in K \cap \text{Lin}(K)^\perp$. Therefore, $K \subset \text{Lin}(K) + K \cap \text{Lin}(K)^\perp$. It follows from (5.8) that

$$K \subset \text{co}(\text{ext}_E(K)) + K^\infty.$$

Thus (5.7) holds as the converse inclusion is obvious. □

Remark. Proposition 5.4 shows that $\text{ext}_E(K)$ is a recession core of K and has the minimality property "up to $\text{Lin}(K)$" in the sense as indicated in Proposition 5.3. In particular, if K contains no lines, then $\text{ext}_E(K) = \text{ext}(K)$ is the least recession core of K.

In terms of recession cores and the BCQs, we now study the global metric subregularity of generalized equation (GEC). Hoffman, in his pioneering work, proved that (GEC) has an error bound (or equivalently, is globally metrically subregular) if $A = X = R^n$ and $F(x) := Qx + R^n_+$ for all $x \in R^n$, where Q is an $m \times n$ matrix. The research on error bounds, especially for inequality systems, has attracted the interest of many researchers and there are a vast number of publications reporting progress in this area. For more details, see [18, 19, 21, 27, 28, 33, 34] and a special issue of Mathematical Programming (Vol. 88, No. 2, 2000). In what follows, we assume that X, Y are general Banach spaces (except when explicitly stated otherwise), F is a closed convex multifunction from X to Y, and A is a closed convex subset of X. In

the case when $A = X$, while the equivalence of (iv) with (v) in the following result is [28, Theorem 3.1], we can now sharpen the result by considering recession cores of S. In what follows, we assume that the solution set S is nonempty.

THEOREM 5.5. *Let C be a recession core of the solution set S of (GEC) and $\tau \in [0, +\infty)$. Then the following statements are equivalent.*

(i) *(GEC) has the strong BCQ at each $x \in C$ with the constant τ.*

(ii) *(GEC) has the strong BCQ at each $x \in S$ with the constant τ.*

(iii) *(GEC) is metrically subregular at each point in C with the constant τ.*

(iv) *(GEC) is metrically subregular at each point in S with the constant τ.*

(v) *(GEC) is globally metrically subregular with the constant τ.*

Proof. (i)⇒(ii) Let $x \in S$. Since C is a recession cone of S, there exist $x_1, \ldots, x_n \in C$, $t_1, \ldots, t_n \in [0, +\infty)$, and $e \in S^\infty$ such that

$$\sum_{i=1}^{n} t_i = 1 \text{ and } x = \sum_{i=1}^{n} t_i x_i + e. \tag{5.9}$$

Let $x^* \in N(S, x) \cap B_{X^*}$. Then $\langle x^*, \sum_{i=1}^n t_i x_i + e\rangle = \max\{\langle x^*, z\rangle : z \in S\}$. Noting that $\sum_{i=1}^n t_i x_i + R_+ e \subset S$, it follows that

$$\langle x^*, e\rangle = 0 \text{ and } \left\langle x^*, \sum_{i=1}^{n} t_i x_i \right\rangle = \max\{\langle x^*, z\rangle : z \in S\}.$$

This implies that for each integer $i \in [1, n]$,

$$\langle x^*, x\rangle = \langle x^*, x_i\rangle = \max\{\langle x^*, z\rangle : z \in S\},$$

and hence $x^* \in N(S, x_i) \cap B_{X^*}$. By (i), one has

$$x^* \in \tau(D^*F(x_i, b)(B_{Y^*}) + N(A, x_i) \cap B_{X^*}), i = 1, \ldots, n.$$

It follows from Lemma 3.1 that $x^* \in \tau(D^*F(x, b)(B_{Y^*}) + N(A, x) \cap B_{X^*})$. Therefore,

$$N(S, x) \cap B_{X^*} \subset \tau(D^*F(x, b)(B_{Y^*}) + N(A, x) \cap B_{X^*}).$$

This shows that (ii) holds. (ii)⇒(i), (iv)⇒(iii), and (v)⇒(iv) are trivial. (iii)⇒(i) and (ii)⇒(iv) are consequences of formula (3.6) in Remark 3.1. The proof of (iv)⇒(v) is similar to that of [28, Theorem 3.1]. Hence, the proof is completed. □

In the special case when $A = X$, $Y = R$, $F(x) = [f(x), +\infty)$ for all $x \in X$, and $b = \inf\{f(x) : x \in X\}$ with f being a proper lower semicontinuous convex function from X to $R \cup \{+\infty\}$, Burke and Deng [3, Theorem 2.3] proved that (GEC) is globally τ-metrically subregular if and only if

$$N(S, z) \cap B_{X^*} \subset \mathrm{cl}^*(\partial f(z)) \quad \forall z \in S.$$

Since $\mathrm{ext}_E(S)$ is a recession core of S if $X = R^n$, the following corollary is a consequence of Theorem 5.5.

COROLLARY 5.6. *Let $X = R^n$. Then (GEC) is globally metrically subregular if and only if there exists $\tau \in (0, +\infty)$ such that (GEC) has the strong BCQ at each generalized extreme point of S with the constant τ.*

Similar to the proof of the equivalent relation (i)⇔(ii) in Theorem 5.5, one can prove the following result.

PROPOSITION 5.7. *Let C be a recession core of S. Then (GEC) has the BCQ at each point in C if and only if (GEC) has the BCQ at each point in S.*

As in the finite dimensional case, let us say that a subset P of X is a polyhedron if there exist $x_n^*, \ldots, x_n^* \in X^*$ and $c_1, \ldots, c_n \in R$ such that

$$P = \{x \in X : \langle x_i^*, x\rangle \le c_i,\ i = 1, \ldots, n\}.$$

It is known that

$$N(P,x) = \left\{ \sum_{i \in I(x)} t_i x_i^* : t_i \ge 0,\ i \in I(x) \right\} \quad \forall x \in P, \tag{5.10}$$

where $I(x) := \{1 \le i \le n : \langle x_i^*, x\rangle = c_i\}$. We say that a multifunction $F : X \to 2^Y$ is polyhedral if its graph is a polyhedron in $X \times Y$. If F is polyhedral, it is easy to verify from (5.10) that

$$N(F^{-1}(b), x) = D^*F(x,b)(Y^*) \quad \forall x \in F^{-1}(b). \tag{5.11}$$

THEOREM 5.8. *Let C be a recession core of the solution set S of (GEC). Suppose that S is a polyhedron in X. Then (GEC) is globally metrically subregular if and only if (GEC) has the BCQ at each point in C.*

Proof. By Theorem 5.5, it suffices to prove the sufficiency. Suppose that (GEC) has the BCQ at each point in C. Then, (GEC) has the BCQ at each point of S (by Proposition 5.7). Since S is a polyhedron, there exist $x_n^*, \ldots, x_n^*$ and $c_1, \ldots, c_n \in R$ such that

$$S = \{x \in X : \langle x_i^*, x\rangle \le c_i,\ i = 1, \ldots, n\}.$$

Let $X_1 := \{x \in X : \langle x_i^*, x\rangle = 0,\ i = 1, \ldots, n\}$. Then X_1 is a closed subspace of X with finite codimension. Thus, there exists a finite dimensional subspace X_2 of X such that $X = X_1 + X_2$ and $X_1 \cap X_2 = \{0\}$. Let

$$P := \{z \in X_2 : \langle x_i^*, z\rangle \le c_i,\ i = 1, \ldots, n\}.$$

It is easy to verify that $S = P + X_1$ and P is a polyhedron containing no lines in X_2. By [31, Theorems 18.5 and 19.1], $P = \mathrm{co}(\mathrm{ext}(P)) + P^\infty$. Hence $S = \mathrm{co}(\mathrm{ext}(P)) + P^\infty + X_1$. Noting that $P^\infty + X_1 \subset (S)^\infty$, it follows that $\mathrm{ext}(P)$ is a recession core of S. Let $e \in \mathrm{ext}(P)$. Then, by (5.10), $N(S,e)$ is a polyhedron in a finite dimensional subspace of X^*. It follows from Proposition 3.4 that there exists $\tau_e \in (0, +\infty)$ such that (GEC) has the strong BCQ with the constant τ_e. Do this for each e in $\mathrm{ext}(P)$ and let $\tau := \max\{\tau_e : e \in \mathrm{ext}(P)\}$. Then $\tau < +\infty$ because $\mathrm{ext}(P)$ is a finite set (cf. [31, Theorem 19.1]). Hence, (GEC) has the strong BCQ at each point of $\mathrm{ext}(P)$ with the constant τ. Since $\mathrm{ext}(P)$ is a recession core of S, it follows from Theorem 5.5 that (GEC) is globally metrically subregular. Hence, the proof is completed. □

In view of the proof of Theorem 5.8, one sees that any polyhedron in a Banach space has a recession core consisting of finitely many elements.

Robinson [30] studied the continuity properties of polyhedral multifunctions. In particular, under the finite dimension assumption, he [30, Corollary] proved that if the graph of F is the union of finitely many polyhedra and $b \in F(X)$, then there exists $\varepsilon, \tau \in [0, +\infty)$ such that

$$d(x, F^{-1}(b)) \le \tau d(b, F(x)) \quad \forall x \in X \text{ with } d(b, F(x)) < \varepsilon.$$

This result can be regarded as a generalization of Hoffman's classical error bound theorem. In the setting of Theorem 5.8, F is not required to be polyhedral but merely the solution set S is required to be polyhedral. When F is a convex polyhedral multifunction and $A = X$, (5.11) implies that generalized equation (GEC) has the BCQ at each $x \in S = F^{-1}(b)$. Hence, in this case, Theorem 5.8 improves the above Robinson's result. But, in the nonconvex case, one cannot use Theorem 5.8 to deduce Robinson's result.

Let $\Gamma(X)$ denote the family of all functions f each of which satisfies the following property: There exist $x_i^* \in X^*$ and continuous convex functions $\phi_i : R \to R$ with $\inf_{t\in R} \phi_i(t) < 0$ $(1 \le i \le n\ )$ such that $f(x) = \max_{1\le i\le n} \phi_i(\langle x_i^*, x\rangle)$ for all $x \in X$. By taking ϕ_i to be affine or quadratic convex functions, we see that $\Gamma(X)$ properly contains the family of all piecewise affine convex functions on X and so the following corollary can be regarded as another generalization of Hoffman's error bound theorem.

COROLLARY 5.9. *Let X be finite dimensional, A be a polyhedron in X, $Y = R$, and $b = 0$. Suppose that there exists $f \in \Gamma(X)$ such that $F(x) = [f(x),\ +\infty)$ for all $x \in X$. Then (GEC) is globally metrically subregular.*

Proof. Let $x_i^* \in X^*$ and $\phi_i : R \to R$ be a continuous convex function with $\inf_{t\in R} \phi_i(t) < 0$ $(1 \le i \le n)$ such that $f(x) = \max_{1\le i\le n} \phi_i(\langle x_i^*, x\rangle)$ for all $x \in X$. Let $P_i = \{x \in X : \langle x_i^*, x\rangle \in \phi_i^{-1}(-\infty, 0]\}$. Since each $\phi_i^{-1}(-\infty, 0]$ is an interval in R, P_i is a polyhedron in X. Hence $S = \bigcap_{i=1}^n A \cap P_i$ is a polyhedron. Let $e \in \mathrm{ext}_E(S)$ and $I(e) := \{1 \le i \le n :\ \phi_i(\langle x_i^*, e\rangle) = f(e)\}$. Then

$$N(S, e) = N(A, e) + \sum_{i=1}^{n} N(P_i, e) = N(A, e) + \sum_{i\in I(e)} N(P_i, e). \tag{5.12}$$

Let $i \in I(e)$. Then, $\phi_i(\langle x_i, e\rangle) = f(e) = 0$ (because $\mathrm{ext}_E(S) \subset \mathrm{bd}(S)$). It follows from $\inf_{t\in R} \phi_i(t) < 0$ that $0 \notin \partial\phi_i(\langle x_i^*, e\rangle)$. By the definition of P_i, there exists $e_i^* \in X^*$ such that $R_+\partial\phi_i(\langle x_i^*, e\rangle)x_i^* \subset N(P_i, e) = R_+e_i^*$. Hence, $N(P_i, e) = R_+\partial\phi_i(\langle x_i^*, e\rangle)x_i^*$ when $x_i^* \ne 0$. Noting that $x_i^* = 0$ implies that $N(P_i, e) = N(X, e) = \{0\}$, it follows from (5.12) that

$$N(S, e) = N(A, e) + \sum_{i\in I(e)} R_+\partial\phi_i(\langle x_i^*, e\rangle)x_i^*. \tag{5.13}$$

On the other hand, by the definition of F, it is easy to verify that

$$D^*F(e, 0)(R) = D^*F(e, 0)(R_+) = R_+\mathrm{co}\left(\bigcup_{i\in I(e)} \partial\phi_i(\langle x_i^*, e\rangle)x_i^*\right).$$

Since $D^*F(e, 0)(R_+)$ is a convex cone, $D^*F(e, 0)(R) = \sum_{i\in I(e)} R_+\partial\phi_i(\langle x_i^*, e\rangle)x_i^*$. This and (5.13) implies that (GEC) has the BCQ at e. It follows from Proposition 5.4 and Theorem 5.8 that (GEC) is globally metrically subregular. The proof is completed. ∎

Let $X = R$ and $f(x) = \max\{(x - 1)^2 - 1,\ (x + 1)^2 - 1\}$ for all $x \in R$. Then $f \in \Gamma(R)$ is not a piecewise affine function and the inequality $f(x) \le 0$ does not satisfy the Slater condition. Examples of this kind provide some interesting cases covered by Corollary 5.9 but neither by Hoffman's error bound theorem nor by the Robinson–Ursescu theorem.

Remark. Let $M : Y \to 2^X$ be a closed convex multifunction and with $(\bar{y}, \bar{x}) \in \mathrm{Gr}(M)$. We say that M is globally calm at $\bar{y}$ over A if there exists $\tau \in (0, +\infty)$ such that

$$d(x, M(\bar{y}) \cap A) \le \tau(\|y - \bar{y}\| + d(x, A)) \quad \forall (x, y) \in \mathrm{Gr}(M).$$

Let C be a recession core of $M(\bar{y}) \cap A$. Theorem 5.5 implies that M is globally calm at $\bar{y}$ over A if and only if there exists $\tau \in (0, +\infty)$ such that

$$N(M(\bar{y}) \cap A, u) \cap B_{X^*} \subset \tau(D^*M^{-1}(u, \bar{y})(B_{Y^*}) + N(A, u) \cap B_{X^*}) \quad \forall u \in C.$$

In the case when $M(\bar{y}) \cap A$ is a polyhedron, Theorem 5.8 implies that M is globally calm at $\bar{y}$ over A if and only if

$$N(M(\bar{y}) \cap A, u) = D^*M^{-1}(u, \bar{y})(Y^*) + N(A, u) \quad \forall u \in C.$$

To end this paper, we provide a procedure to find the generalized extreme points of a polyhedron in a finite dimensional space. Let $a_1, \ldots, a_m \in R^n$, $c_1, \ldots, c_m \in R$, and let P denote the polyhedron determined by a_i and c_i $(i = 1, \ldots, m)$, that is,

$$P = \{x \in R^n : \langle a_i, x\rangle \le c_i,\ i = 1, \ldots, m\}.$$

For convenience, let $I := \{1, \ldots, m\}$ and $I(x) = \{i \in I : \langle a_i, x\rangle = c_i\}$ for $x \in P$. Let $\mathcal{M}(I)$ denote the family of all subsets D of I with the property that $\{a_i : i \in D\}$ is a maximal linearly independent subset of $\{a_i : i \in I\}$. Thus elements of $\mathcal{M}(I)$ can be obtained by the Gram–Schmidt process. For each $D \in \mathcal{M}(I)$, the linear equation system

$$\sum_{j \in D} \langle a_i, a_j\rangle t_j = c_i \quad \forall i \in D$$

has a unique solution which will be denoted by $(\bar{t}_j)_{j \in D}$; we shall also write e_D for $\sum_{j \in D} \bar{t}_j a_j$. Let

$$\mathcal{E}(I) = \{D \in \mathcal{M}(I) : \langle a_i, e_D\rangle \le c_i,\ i \in I \setminus D\}.$$

THEOREM 5.10. $\mathrm{ext}_E(P) = \{e_D : D \in \mathcal{E}(I)\}$.

Proof. Note that $\mathrm{Lin}(P) = \{x \in R^n : \langle a_i, x\rangle = 0,\ i \in I\}$. Hence,

$$\mathrm{Lin}(P)^\perp = \mathrm{span}\{a_i : i \in I\} = \mathrm{span}\{a_i : i \in D\} \quad \forall D \in \mathcal{M}(I), \tag{5.14}$$

where span$A$ denotes the linear hull of A. Let $e \in \mathrm{ext}_E(P)$ and pick a $D_0 \subset I(e)$ such that $\{a_i : i \in D_0\}$ is a maximal linearly independent subset of $\{a_i : i \in I(e)\}$. We claim that $D_0 \in \mathcal{M}(I)$. Indeed, if this is not the case, then $\mathrm{span}\{a_i : i \in I(e)\}$ is a proper subspace of $\mathrm{span}\{a_i : i \in I\}$. It follows from the first equality of (5.14) that there exists $h \in \mathrm{Lin}(P)^\perp \setminus \{0\}$ such that $\langle a_i, h\rangle = 0$ for all $i \in I(e)$. Since $\langle a_i, e\rangle < c_i$ for all $i \in I \setminus I(e)$, there exists $\varepsilon > 0$ small enough such that $e \pm \varepsilon h \in P$. Since $e = \frac{e+\varepsilon h+(e-\varepsilon h)}{2}$, it follows from (5.4) that $2\varepsilon h \in \mathrm{Lin}(P)$. This contradicts $h \in \mathrm{Lin}(P)^\perp \setminus \{0\}$. Hence $D_0 \in \mathcal{M}(I)$. Noting that $e \in \mathrm{Lin}(P)^\perp$ (by Proposition 5.2), it follows from (5.14) that there exists $(\bar{t}_j)_{j \in D_0} \in R^{|D_0|}$ such that $e = \sum_{j \in D_0} \bar{t}_j a_j$, where $|D_0|$ denotes the number of elements of D_0. It follows from $e \in P$ and $D_0 \subset I(e)$ that $D_0 \in \mathcal{E}(I)$ and $e = e_{D_0}$. Therefore, $\mathrm{ext}_E(P) \subset \{e_D : D \in \mathcal{E}(I)\}$. It remains to be seen whether $\{e_D : D \in \mathcal{E}(I)\} \subset \mathrm{ext}_E(P)$. To do this, let $D \in \mathcal{E}(I)$. Then

$e_D \in P \cap \mathrm{Lin}(P)^{\perp}$ (by (5.14) and the definition of e_D). Let $x_1, x_1 \in P$ satisfy $e_D = \frac{x_1+x_2}{2}$. It follows that $\langle a_i, x_1\rangle = \langle a_i, x_2\rangle = c_i$ for all $i \in D$ and so $\langle a_i, x_1 - x_2\rangle = 0$ for all $i \in D$. Since $\{a_i : i \in D\}$ is a maximal linearly independent subset of $\{a_i : i \in I\}$, $\langle a_i, x_1 - x_2\rangle = 0$ for all $i \in I$. Hence $x_1 - x_2 \in \mathrm{Lin}(P)$. This shows that $e_D \in \mathrm{ext}_E(P)$. Hence, the proof is completed. □

Acknowledgments. The authors would like to express their sincere thanks to Professor Boris Mordukhovich for stimulating discussions and suggestions. They also thank the associate editor and the referees for their helpful comments and for pointing out references [16, 17, 23].

REFERENCES

[1] J.-P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhäuser Boston, Boston, 1990.
[2] J. F. Bonnans and A. Shapiro, *Perturbation Analysis of Optimization Problems*, Springer-Verlag, New York, 2000.
[3] J. V. Burke and S. Deng, *Weak sharp minima revisited, Part I: Basic theory*, Control Cybernetics, 31 (2002), pp. 399–469.
[4] J. V. Burke, M. C. Ferris, and M. Qian, *On the Clarke subdifferential of the distance function of a closed set*, J. Math. Anal. Appl., 166 (1992), pp. 199–213.
[5] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley & Sons, New York, 1983.
[6] A. L. Dontchev, A. S. Lewis, and R. T. Rockafellar, *The radius of metric regularity*, Trans. Amer. Math. Soc., 355 (2003), pp. 493–517.
[7] A. L. Dontchev and R. T. Rockafellar, *Regularity and conditioning of solution mappings in variational analysis*, Set-Valued Anal., 12 (2004), pp. 79–109.
[8] R. Henrion and A. Jourani, *Subdifferential conditions for calmness of convex constraints*, SIAM J. Optim., 13 (2002), pp. 520–534.
[9] R. Henrion, A. Jourani, and J. Outrata, *On the calmness of a class of multifunctions*, SIAM J. Optim., 13 (2002), pp. 603–618.
[10] R. Henrion and J. Outrata, *Calmness of constraint systems with applications*, Math. Program., 104 (2005), pp. 437–464.
[11] J.-B. Hiriart-Urruty and C. Lemarechal, *Convex Analysis and Minimization Algorithms* I. Fundamentals, Springer-Verlag, Berlin, 1993.
[12] H. Hu, *Characterizations of the strong basic constraint qualification*, Math. Oper. Res., 30 (2005), pp. 956–965.
[13] A. D. Ioffe, *Metric regularity and subdifferential calculus*, Russian Math. Surveys, 55 (2000), pp. 501–558.
[14] G. Jameson, *Ordered Linear Spaces*, Lecture Notes in Mathematics 141, Springer-Verlag, Berlin, 1970.
[15] D. Klatte and B. Kummer, *Nonsmooth Equations in Optimization*, in Regularity, Calculus, Methods, and Applications, Nonconvex Optimization and its Application 60, Kluwer Academic Publishers, Dordrecht, 2002.
[16] D. Klatte and B. Kummer, *Constrained minima and Lipschitzian penalties in metric spaces*, SIAM J. Optim., 13 (2002), pp. 619–633.
[17] A. Y. Kruger, *Covering theorem for set-valued mappings*, Optimization, 19 (1988), pp. 763–780.
[18] A. Lewis and J. S. Pang, *Error bounds for convex inequality systems*, in Generalized Convexity, Generalized Monotonicity: Recent Results, Proceedings of the Fifth Symposium on Generalized Convexity, Luminy, France, 1996, J.-P. Crouzeix, J.-E. Martinez-Legaz, and M. Volle, eds., Kluwer Academic Publishers, Dordrecht, 1998, pp. 75–100.
[19] W. Li, *Abadie's constraint qualification, metric regularity, and error bounds for differentiable convex inequalities*, SIAM J. Optim., 7 (1997), pp. 966–978.
[20] W. Li, C. Nahak, and I. Singer, *Constraint qualifications for semi-infinite systems of convex inequalities*, SIAM J. Optim., 11 (2000), pp. 31–52.
[21] W. Li and I. Singer, *Global error bounds for convex multifunctions and applications*, Math. Oper. Res., 23 (1998), pp. 443–462.
[22] B. S. Mordukhovich, *Complete characterization of openness, metric regularity, and Lipschitzian properties of multifunctions*, Trans. Amer. Math. Soc., 340 (1993), pp. 1–35.
[23] B. S. Mordulhovich, *Coderivatives of set-valued mappings: Calculus and applications*, Nonlinear Anal., 30 (1997), pp. 3059–3070.

[24] B. S. Mordukhovich, *Variational Analysis and Generalized Differentiation* I*, Basic Theory*, Springer-Verlag, Berlin, 2006.
[25] B. S. Mordukhovich, *Variational Analysis and Generalized Differentiation* II*, Applications*, Springer-Verlag, Berlin, 2006.
[26] B. S. Mordukhovich and Y. Shao, *Nonconvex differential calculus for infinite-dimensional multifunctions*, Set-Valued Anal., 4 (1996), pp. 205–236.
[27] K. F. Ng and W. H. Yang, *Error bounds for abstract linear inequality systems*, SIAM J. Optim., 13 (2002), pp. 24–43.
[28] K. F. Ng and X. Y. Zheng, *Characterizations of error bounds for convex multifunctions on Banach spaces*, Math. Oper. Res., 29 (2004), pp. 45–63.
[29] S. M. Robinson, *Normed convex processes*, Trans. Amer. Math. Soc., 174 (1972), pp. 127–140.
[30] S. M. Robinson, *Some continuity properties of polyhedral multifunctions*, Math. Programing Stud., 14 (1981), pp. 206–214.
[31] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[32] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
[33] C. Ursescu, *Multifunctions with convex closed graph*, Czechoslovak Math. J., 25 (1975), pp. 438–441.
[34] C. Zalinescu, *A nonlinear extension of Hoffman's error bounds for linear inequalities*, Math. Oper. Res., 28 (2003), pp. 524–532.
[35] C. Zalinescu, *Weak sharp minima, well-behaving functions, and global error bounds for convex inequalities in Banach spaces*, in Proceedings of the 12th Baical International Conference on Optimization Methods and their Applications, Irkutsk, Russia, 2001, pp. 272–284.
[36] X. Y. Zheng and K. F. Ng, *Metric regularity and constraint qualifications for convex inequalities on Banach spaces*, SIAM J. Optim., 14 (2003), pp. 757–772.

SIAM J. OPTIM.
Vol. 18, No. 2, pp. 461–481

SOME GLOBAL UNIQUENESS AND SOLVABILITY RESULTS FOR LINEAR COMPLEMENTARITY PROBLEMS OVER SYMMETRIC CONES*

M. SEETHARAMA GOWDA† AND R. SZNAJDER‡

Abstract. This article deals with linear complementarity problems over symmetric cones. Our objective here is to characterize global uniqueness and solvability properties for linear transformations that leave the symmetric cone invariant. Specifically, we show that, for algebra automorphisms on the Lorentz space $\mathcal{L}^n$ and for quadratic representations on any Euclidean Jordan algebra, global uniqueness, global solvability, and the $\mathbf{R}_0$-properties are equivalent. We also show that for Lyapunov-like transformations, the global uniqueness property is equivalent to the transformation being positive stable and positive semidefinite.

Key words. Euclidean Jordan algebra, symmetric cone, algebra/cone automorphism, $\mathbf{R}_0$-property, **Q**-property, **GUS**-property

AMS subject classifications. Primary, 90C33, 17C55; Secondary, 15A48, 15A57

DOI. 10.1137/060653640

1. Introduction.

Given a finite dimensional real inner product space H, a closed convex set K in H, a continuous function $f : K \to H$, and a vector $q \in H$, the *variational inequality problem* $\mathrm{VI}(f, K, q)$ is to find an $x^* \in K$ such that

$$\langle f(x^*) + q, x - x^* \rangle \geq 0 \ \forall \ x \in K.$$

There is an extensive literature associated with this problem covering theory, applications, and computation of solutions; see, e.g., [7]. When K is a closed convex cone, this problem reduces to the cone complementarity problem $\mathrm{CP}(f, K, q)$, which further reduces to the linear complementarity problem $\mathrm{LCP}(f, K, q)$ when f is linear. In particular, when $H = R^n$ (with the usual inner product), $f\,(= M)$ is linear, and $K = R^n_+$, this reduces to the *standard linear complementarity problem* $\mathrm{LCP}(M, R^n_+, q)$ [3].

An unsolved problem in the variational inequality theory is the characterization of the global uniqueness property: Given H and K, find a necessary and sufficient condition on f so that for all $q \in H$, $\mathrm{VI}(f, K, q)$ has a unique solution. This is related to the question of global invertibility of the normal map $F(x) := f(\Pi_K(x)) + x - \Pi_K(x)$ on H; see [7]. When K is polyhedral and f is linear, there is a well-known result of Robinson [20] that describes the invertibility of this map in terms of the determinants of a certain collection of matrices. This result, when specialized to the standard linear complementarity problem, says that for a square real matrix M, the standard linear complementarity problem $\mathrm{LCP}(M, R^n_+, q)$ has a unique solution for all q if and only if M is a **P**-matrix (which means that all principal minors of M are positive). The result of Robinson motivated researchers to consider the (more general) question of

*Received by the editors May 9, 2006; accepted for publication (in revised form) January 15, 2007; published electronically May 22, 2007.
http://www.siam.org/journals/siopt/x-x/65943.html
†Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD 21250 (gowda@math.umbc.edu, http://www.math.umbc.edu/~gowda).
‡Department of Mathematics, Bowie State University, Bowie, MD 20715 (rsznajder@bowiestate.edu).

global invertibility of piecewise affine functions; see [24], [8] for necessary and sufficient conditions.

Moving away from the polyhedral settings (where the underlying cone is the nonnegative orthant in R^n) and inspired by the recent interest in conic programming, various researchers have started looking at cone linear complementarity problems, particularly on semidefinite and second-order cones, and more generally on symmetric cones. While symmetric cones are, in general, nonpolyhedral, they have a lot of structure. In spite of this, even in this special case (of a linear transformation on a symmetric cone) the global uniqueness problem remains unsolved. At present, various authors have studied this problem by restricting the symmetric cone and the linear transformation to specific classes. Here are some such results:

(1) Given a matrix $A \in R^{n\times n}$, consider the Lyapunov transformation L_A defined on the space $\mathcal{S}^n$ of all $n \times n$ real symmetric matrices by

$$L_A(X) := AX + XA^T.$$

Then it has been shown by Gowda and Song [10] that L_A has the global uniqueness property on the semidefinite cone $\mathcal{S}^n_+$ (i.e., for all $q \in \mathcal{S}^n$, LCP$(L_A, \mathcal{S}^n_+, q)$ has a unique solution) if and only if A is both positive stable (i.e., all of its eigenvalues have positive real parts) and positive semidefinite.

(2) Given a matrix $A \in R^{n\times n}$, define the multiplication transformation M_A defined on $\mathcal{S}^n$ by

$$M_A(X) := AXA^T.$$

Then it has been shown by Bhimasankaram et al. [2] and Gowda, Song, and Ravindran [11] that M_A has the global uniqueness property on the semidefinite cone if and only if $\pm A$ is positive definite.

(3) On the Lorentz space $\mathcal{L}^n$ (see section 2 for the definition), consider the quadratic representation P_a of an element $a \in \mathcal{L}^n$:

$$P_a(x) := 2a \circ (a \circ x) - a^2 \circ x.$$

In this setting, Malik and Mohan [17] have shown that P_a has the global uniqueness property on the Lorentz cone if and only if $\pm a$ is in the interior of that cone.

Another issue in the variational inequality/complementarity theory is the global solvability: Given H and K, find a necessary and sufficient condition on f so that VI(f, K, q) has a solution for all $q \in H$. We note that this remains unsolved even in the setting of the standard linear complementarity problem. So, as in the uniqueness issue, one has to work within a class of cones/transformations to get meaningful results. Our motivation for this part comes from the following:

(a) For a nonnegative matrix M, Murty [18] has shown that M has the global solvability property with respect to the nonnegative orthant R^n_+ (i.e., LCP(M, R^n_+, q) has a solution for all $q \in R^n$) if and only if the diagonal of M is positive.

(b) The Lyapunov transformation L_A (defined earlier) has the global solvability property with respect to $\mathcal{S}^n_+$ if and only if A is positive stable [10].

(c) For matrix $A \in R^{n\times n}$, consider the Stein transformation S_A defined on the space $\mathcal{S}^n$ by

$$S_A(X) := X - AXA^T.$$

Then S_A has the global solvability property on $\mathcal{S}^n_+$ if and only if A is Schur stable (that is, all eigenvalues of A lie in the open unit disk of the complex plane) [9].

(d) Given a real matrix A, consider the multiplication transformation M_A (defined earlier) on $\mathcal{S}^n$. In [21] [19], Sampangi Raman, has shown that when A is symmetric M_A has the global solvability property with respect to the semidefinite cone in $\mathcal{S}^n_+$ if and only if $\pm A$ is positive definite.

(e) In [17], Malik and Mohan have shown that P_a on $\mathcal{L}^n$ has the global solvability property with respect to the Lorentz cone if and only if $\pm a$ is in the interior of the Lorentz cone.

In keeping with the above global uniqueness/solvability issues, we consider linear complementarity problems over symmetric cones in Euclidean Jordan algebras. With the observation (elaborated in various sections of the paper) that all of the transformations considered in items (1)–(3) and (a)–(e) either leave the symmetric cone invariant or are related to one such, we characterize global uniqueness/solvability properties for algebra automorphisms, quadratic representations, and Lyapunov-like transformations.

Here is a brief description/outline of our paper. Let V be a Euclidean Jordan algebra V with the corresponding symmetric cone K. For a linear transformation $L : V \to V$ and a $q \in V$, we define the (cone) linear complementarity problem $\mathrm{LCP}(L, K, q)$ as the problem of finding $x \in V$ such that

$$x \in K,\ L(x) + q \in K, \quad \text{and} \quad \langle L(x) + q, x\rangle = 0.$$

Given L, we consider the following statements:

(α) For all $q \in V$, $\mathrm{LCP}(L, K, q)$ has a unique solution.

(β) For all $q \in V$, $\mathrm{LCP}(L, K, q)$ has a solution.

(γ) $\mathrm{LCP}(L, K, 0)$ has zero as the only solution.

• In section 4, we show that for algebra automorphisms on $\mathcal{L}^n$ (these are invertible linear transformations satisfying $L(x \circ y) = L(x) \circ L(y) \quad \forall\, x, y$) the above three properties are equivalent; this result may be regarded as an analog of item (2) above for algebra automorphisms on $\mathcal{L}^n$.

• In section 5, we show that the above three properties are equivalent for any quadratic representation (given by $P_a(x) = 2a \circ (a \circ x) - a^2 \circ x$) on any Euclidean Jordan algebra, thereby extending Malik–Mohan's result (items (3) and (e) above) to arbitrary Euclidean Jordan algebras.

• In section 6, we show that if L is Lyapunov-like, that is, if it satisfies the condition

$$x, y \in K, \text{ and } \langle x, y\rangle = 0 \Rightarrow \langle L(x), y\rangle = 0,$$

then the global uniqueness property (α) holds if and only if L is positive stable and positive semidefinite, thereby extending the result of item (1) above to general Euclidean Jordan algebras.

2. Preliminaries.

2.1. Euclidean Jordan algebras. In this paper we deal with Euclidean Jordan algebras. For the sake of completeness, we provide a short introduction (as in [22]); for full details, see [6].

A *Euclidean Jordan algebra* is a triple $(V, \circ, \langle\cdot,\cdot\rangle)$, where $(V, \langle\cdot,\cdot\rangle)$ is a finite dimensional inner product space over R and $(x, y) \mapsto x \circ y : V \times V \to V$ is a bilinear mapping satisfying the following conditions:

(i) $x \circ y = y \circ x$ for all $x, y \in V$;

(ii) $x \circ (x^2 \circ y) = x^2 \circ (x \circ y)$ for all $x, y \in V$, where $x^2 := x \circ x$; and

(iii) $\langle x \circ y, z\rangle = \langle y, x \circ z\rangle$ for all $x, y, z \in V$.

We also assume that there is an element $e \in V$ (called the *unit* element) such that $x \circ e = x$ for all $x \in V$.

In a Euclidean Jordan algebra V, the set of squares

$$K := \{x \circ x : x \in V\}$$

is a *symmetric cone* (see p. 46, Faraut and Korányi [6]). This means that K is a self-dual closed convex cone (i.e., $K = K^* := \{x \in V : \langle x, y\rangle \geq 0 \; \forall y \in K\}$) and for any two elements $x, y \in K^\circ (= \text{interior}(K))$, there exists an invertible linear transformation $\Gamma : V \to V$ such that $\Gamma(K) = K$ and $\Gamma(x) = y$. We use the notation

$$x \geq 0 \quad \text{and} \quad x > 0$$

when $x \in K$ and $x \in K^\circ$, respectively.

An element $c \in V$ is an *idempotent* if $c^2 = c$; it is a *primitive idempotent* if it is nonzero and cannot be written as a sum of two nonzero idempotents.

We say that a finite set $\{e_1, e_2, \ldots, e_m\}$ of idempotents in V is a *a complete system of orthogonal idempotents* if

$$e_i \circ e_j = 0 \text{ for } i \neq j, \text{ and } \sum_1^m e_i = e.$$

(Note that $\langle e_i, e_j\rangle = \langle e_i \circ e_j, e\rangle = 0$ whenever $i \neq j$.) Further, if each e_i is also primitive, we say that the system is a *Jordan frame*.

THEOREM 2.1 (the spectral decomposition theorem) (Thms. III.1.1 and III.1.2, Faraut and Korányi [6]). *Let V be a Euclidean Jordan algebra. Then there is a number r (called the rank of V) such that for every $x \in V$, there exists a Jordan frame $\{e_1, \ldots, e_r\}$ and real numbers $\lambda_1, \ldots, \lambda_r$, with*

$$x = \lambda_1 e_1 + \cdots + \lambda_r e_r. \tag{2.1}$$

Also, for each $x \in V$, there exists a unique set of distinct real numbers $\{\mu_1, \mu_2, \ldots, \mu_k\}$ and a unique complete system of orthogonal idempotents $\{f_1, f_2, \ldots, f_k\}$ such that

$$x = \mu_1 f_1 + \mu_2 f_2 + \cdots + \mu_k f_k.$$

For an x given by (2.1), the numbers $\lambda_1, \lambda_2, \ldots, \lambda_r$ are called the eigenvalues of x. We say that x is *invertible* if every λ_i is nonzero. Corresponding to (2.1), we define

$$\text{trace}(x) := \lambda_1 + \lambda_2 + \cdots + \lambda_r.$$

In addition, when $x \geq 0$ (or, equivalently, every λ_i is nonnegative), we define

$$\sqrt{x} := \sqrt{\lambda_1} e_1 + \cdots + \sqrt{\lambda_r} e_r.$$

In a Euclidean Jordan algebra V, for a given $x \in V$, we define the corresponding *Lyapunov transformation* $L_x : V \to V$ by

$$L_x(z) = x \circ z.$$

We say that elements x and y *operator commute* if $L_x L_y = L_y L_x$. It is known that x and y operator commute if and only if x and y have their spectral decompositions with respect to a common Jordan frame (Lem. X.2.2, Faraut and Korányi [6]).

Here are some standard examples.

Example 1. R^n is a Euclidean Jordan algebra with the inner product and the Jordan product defined, respectively, by

$$\langle x, y\rangle = \sum_{i=1}^{n} x_i y_i \quad \text{and} \quad x \circ y := (x_i y_i).$$

Here R^n_+ is the corresponding symmetric cone.

Example 2. $\mathcal{S}^n$, the set of all $n \times n$ real symmetric matrices, is a Euclidean Jordan algebra with the inner and Jordan products given by

$$\langle X, Y\rangle := \text{trace}(XY) \quad \text{and} \quad X \circ Y := \frac{1}{2}(XY + YX).$$

In this setting, the symmetric cone $\mathcal{S}^n_+$ is the set of all positive semidefinite matrices in $\mathcal{S}^n$. Also, X and Y operator commute if and only if $XY = YX$.

Example 3. Consider R^n $(n > 1)$ where any element x is written as

$$x = \begin{bmatrix} x_0 \\ \overline{x} \end{bmatrix},$$

with $x_0 \in R$ and $\overline{x} \in R^{n-1}$. The inner product in R^n is the usual inner product. The Jordan product $x \circ y$ in R^n is defined by

$$x \circ y = \begin{bmatrix} x_0 \\ \overline{x} \end{bmatrix} \circ \begin{bmatrix} y_0 \\ \overline{y} \end{bmatrix} := \begin{bmatrix} \langle x, y\rangle \\ x_0\overline{y} + y_0\overline{x} \end{bmatrix}.$$

We shall denote this Euclidean Jordan algebra $(R^n, \circ, \langle\cdot,\cdot\rangle)$ by $\mathcal{L}^n$. In this algebra, the cone of squares, denoted by $\mathcal{L}^n_+$, is called the *Lorentz cone* (or the second-order cone). It is given by

$$\mathcal{L}^n_+ = \{x : x_0 \geq ||\overline{x}||\},$$

in which case interior$(\mathcal{L}^n_+) = \{x : x_0 > ||\overline{x}||\}$.

We note the spectral decomposition of any x with $\overline{x} \neq 0$:

$$x = \lambda_1 e_1 + \lambda_2 e_2,$$

where

$$\lambda_1 := x_0 + ||\overline{x}||, \ \lambda_2 := x_0 - ||\overline{x}||,$$

and

$$e_1 := \frac{1}{2}\begin{bmatrix} 1 \\ \frac{\overline{x}}{||\overline{x}||} \end{bmatrix}, \quad e_2 := \frac{1}{2}\begin{bmatrix} 1 \\ -\frac{\overline{x}}{||\overline{x}||} \end{bmatrix}.$$

In $\mathcal{L}^n$, elements x and y operator commute if and only if either $\overline{x}$ is a multiple of $\overline{y}$ or $\overline{y}$ is a multiple of $\overline{x}$.

We recall the following from Gowda, Sznajder, and Tao [12] (with the notation $x \geq 0$ when $x \in K$).

PROPOSITION 2.2. *For x, $y \in V$, the following conditions are equivalent:*

1. *$x \geq 0$, $y \geq 0$, and $\langle x, y\rangle = 0$.*
2. *$x \geq 0$, $y \geq 0$, and $x \circ y = 0$.*

In each case, elements x and y operator commute.

The Peirce decomposition. Fix a Jordan frame $\{e_1, e_2, \ldots, e_r\}$ in a Euclidean Jordan algebra V. For $i, j \in \{1, 2, \ldots, r\}$, define the eigenspaces

$$V_{ii} := \{x \in V : x \circ e_i = x\} = R\, e_i,$$

and when $i \neq j$,

$$V_{ij} := \left\{x \in V : x \circ e_i = \frac{1}{2}x = x \circ e_j\right\}.$$

Then we have the following result.

THEOREM 2.3 (Thm. IV.2.1, Faraut and Korányi [6]). *The space V is the orthogonal direct sum of spaces V_{ij} $(i \leq j)$. Furthermore,*

$$\begin{aligned} &V_{ij} \circ V_{ij} \subset V_{ii} + V_{jj}, \\ &V_{ij} \circ V_{jk} \subset V_{ik} \text{ if } i \neq k, \\ &V_{ij} \circ V_{kl} = \{0\} \text{ if } \{i, j\} \cap \{k, l\} = \emptyset. \end{aligned}$$

Thus, given any Jordan frame $\{e_1, e_2, \ldots, e_r\}$, we have the *Peirce decomposition* of $x \in V$:

$$x = \sum_{i=1}^{r} x_{ii} + \sum_{i<j} x_{ij} = \sum_{i=1}^{r} x_i e_i + \sum_{i<j} x_{ij},$$

where $x_i \in R$ and $x_{ij} \in V_{ij}$.

A Euclidean Jordan algebra is said to be *simple* if it is not a direct sum of two Euclidean Jordan algebras. The classification theorem (Chap. V, Faraut and Korányi [6]) says that every simple Euclidean Jordan algebra is isomorphic to one of the following:

(1) the algebra $\mathcal{S}^n$ of $n \times n$ real symmetric matrices,
(2) the algebra $\mathcal{L}^n$,
(3) the algebra $\mathcal{H}_n$ of all $n \times n$ complex Hermitian matrices with a trace inner product and $X \circ Y = \frac{1}{2}(XY + YX)$,
(4) the algebra $\mathcal{Q}_n$ of all $n \times n$ quaternion Hermitian matrices with a (real) trace inner product and $X \circ Y = \frac{1}{2}(XY + YX)$, and
(5) the algebra $\mathcal{O}_3$ of all 3×3 octonion Hermitian matrices with a (real) trace inner product and $X \circ Y = \frac{1}{2}(XY + YX)$.

The following result characterizes all Euclidean Jordan algebras.

THEOREM 2.4 (Props. III.4.4 and III.4.5 and Thm. V.3.7, Faraut and Korányi [6]). *Any Euclidean Jordan algebra is, in a unique way, a direct sum of simple Euclidean Jordan algebras. Moreover, the symmetric cone in a given Euclidean Jordan algebra is, in a unique way, a direct sum of symmetric cones in the constituent simple Euclidean Jordan algebras.*

2.2. Complementarity concepts. Let V be a Euclidean Jordan algebra with the corresponding symmetric cone K. Recall that for a linear transformation $L : V \to V$ and $q \in V$, the *linear complementarity problem* $\mathrm{LCP}(L, K, q)$ is to find an $x \in V$ such that

$$x \in K,\ L(x) + q \in K, \quad \text{and} \quad \langle L(x) + q, x\rangle = 0.$$

As mentioned previously, this is a particular instance of a variational inequality problem. The standard linear complementarity problem (over the nonnegative orthant

in R^n), the semidefinite linear complementarity problem, and the Lorentz cone (also known as the second-order cone) linear complementarity problem are some of the special cases and have been well studied in the literature. Given L on V, we say that L has the

(i) positive definite property if $\langle L(x), x\rangle > 0$ for all $x \neq 0$;
(ii) **GUS** (globally uniquely solvable)-property if for all $q \in V$, LCP(L, K, q) has a unique solution;
(iii) **P**-property if

$$x \text{ and } L(x) \text{ operator commute and } x \circ L(x) \leq 0 \Rightarrow x = 0;$$

(iv) $\mathbf{R}_0$-property if zero is the only solution of LCP$(L, K, 0)$;
(v) **Q**-property if for all $q \in V$, LCP(L, K, q) has a solution;
(vi) **S**-property if there exists a $d > 0$ such that $L(d) > 0$.

Henceforth, we will use **P**, $\mathbf{R}_0$, **Q**, **S**, etc., to denote the set of maps L that satisfy the respective property.

The above properties have been well studied. In particular (see Thms. 17, 14, and 12, [12]), we always have the implications (i) $\Rightarrow$ (ii) $\Rightarrow$ (iii) $\Rightarrow$ (iv). That (v) $\Rightarrow$ (vi) follows from perturbing a solution of LCP$(L, K, -e)$, where e is the unit element of V.

The following well-known result shows that under an additional assumption, (iv) $\Rightarrow$ (v).

THEOREM 2.5 (Karamardian [15]). *Suppose that $L : V \to V$ is a linear transformation such that for some $d > 0$, zero is the only solution of the problems* LCP$(L, K, 0)$ *and* LCP(L, K, d). *Then L has the* **Q***-property with respect to K.*

2.3. Automorphisms. Let V be a Euclidean Jordan algebra and K be the corresponding cone of squares. We consider the following sets of transformations:

- $Aut(V)$—set of all invertible linear transformations $L : V \to V$ such that

$$L(x \circ y) = L(x) \circ L(y) \ \forall x, y \in V,$$

- $Aut(K)$—set of all (invertible) linear transformations $L : V \to V$ such that $L(K) = K$,
- $\overline{Aut(K)}$—closure of $Aut(K)$ (with respect to the operator norm), and
- $\Pi(K)$—set of all linear transformations $L : V \to V$ such that $L(K) \subseteq K$.

We note that

$$Aut(V) \subseteq Aut(K) \subseteq \overline{Aut(K)} \subseteq \Pi(K).$$

Also, if V is simple or if the inner product in V is given by $\langle x, y\rangle = c \text{ trace}(x \circ y)$ (for some fixed c), then every L in $Aut(V)$ is orthogonal (that is, it preserves the inner product); see p. 57, [6].

3. Cone invariant transformations. Recall that $\Pi(K)$ is the set of all linear transformations on V that leave K invariant; for properties, see [1]. We begin by describing some complementarity properties of $\Pi(K)$ and $\overline{Aut(K)}$.

PROPOSITION 3.1. *For $L \in \Pi(K)$, the following are equivalent:*

(a) *L has the $\mathbf{R}_0$-property.*
(b) *For any primitive idempotent $u \in V$, $\langle L(u), u\rangle > 0$.*
(c) *L is strictly copositive on K; i.e., $\langle L(x), x\rangle > 0$ for all $0 \neq x \geq 0$.*

In particular, if L has the $\mathbf{R}_0$-property, then it has the **Q***-property.*

Proof. From $L \in \Pi(K)$, we see that L is copositive on K; that is, $\langle L(x), x\rangle \geq 0$ for all $x \geq 0$. Now assume (a). For any primitive idempotent u, we have $L(u) \in K$, and so $\langle L(u), u\rangle \geq 0$. If $\langle L(u), u\rangle = 0$, then u will be a nonzero solution of LCP$(L, K, 0)$ contradicting condition (a). Hence (b) holds. Now suppose (b) holds. We know that L is copositive on K. Suppose, if possible, $\langle L(x), x\rangle = 0$ for some nonzero $x \in K$. Let $x = \sum \lambda_i e_i$ be the spectral decomposition of x with some eigenvalue, say, $\lambda_k \neq 0$. As $\lambda_i \geq 0$ for all i, $\langle L(x), x\rangle = \sum_{i,j} \lambda_i \lambda_j \langle L(e_i), e_j\rangle \geq \lambda_k^2 \langle L(e_k), e_k\rangle > 0$ by condition (b). This is a contradiction. Hence (c) holds. Finally, the implication (c) $\Rightarrow$ (a) is obvious.

Now assume that (a) holds. Then L is strictly copositive on K, and so the problems LCP$(L, K, 0)$ and LCP(L, K, e) have zero as the only solution. By Karamardian's theorem, LCP(L, K, q) has a solution for all $q \in V$. Thus L has the **Q**-property. ∎

PROPOSITION 3.2. *If $L \in \overline{Aut(K)}$ is invertible, then $L \in Aut(K)$.*

Proof. Let $L_k \in Aut(K)$ such that $L_k \to L$ (with respect to the operator norm) on V with L invertible. From $L_k(K) \subseteq K$, we get $L(K) \subseteq K$. Also, $L_k^{-1} \to L^{-1}$. From $(L_k)^{-1}(K) \subseteq K$, we get $L^{-1}(K) \subseteq K$. Thus we have $L(K) = K$. ∎

PROPOSITION 3.3. $\overline{Aut(K)} \cap \mathbf{S} = Aut(K)$.

Proof. Recall that $L \in \mathbf{S}$ if there exists a $p > 0$ such that $L(p) > 0$. Then, clearly, $Aut(K)$ is contained in $\overline{Aut(K)} \cap \mathbf{S}$. Now suppose that $L \in \overline{Aut(K)} \cap \mathbf{S}$. Let $L = \lim L_k$, where $L_k \in Aut(K)$. As $L_k^T \in Aut(K)$ (See Prop. I.1.7, [6]) and $L^T = \lim L_k^T$, we have $L^T \in \overline{Aut(K)}$. We show that L^T is invertible (or, equivalently, L is invertible) and then conclude (thanks to the previous proposition) that $L \in Aut(K)$. Now to show that L^T is invertible, we show that for each $d > 0$ there is an $x \in K$ such that $L^T(x) = d$. This then shows that the range of L^T contains the open set K°, thus proving the invertibility of L^T. Now fix a $d > 0$. Since $L_k^T \in Aut(K)$ for each k, there exists a sequence $x_k \in K$ such that $L_k^T(x_k) = d$ for all k. Assume, if possible, that the sequence x_k is unbounded, say, $||x_k|| \to \infty$. As x_k is nonzero, we can let $\frac{x_k}{||x_k||} \to y \in K$ with $L^T(y) = 0$. Now because L has the **S**-property, there exists a $p > 0$ such that $L(p) > 0$. If u is a suitable multiple of p, then $u \geq 0$ and $v = L(u) - d \geq 0$. Then

$$0 \leq \langle y, v\rangle = \langle y, L(u) - d\rangle = \langle L^T(y), u\rangle - \langle y, d\rangle = -\langle y, d\rangle \leq 0.$$

Thus, $\langle y, d\rangle = 0$. Since $y \geq 0$ and $d > 0$, we must have $y = 0$, which is a contradiction. Hence the sequence x_k is bounded. Letting $x_k \to x \in K$, we have $L^T(x) = d$. ∎

COROLLARY 3.4.

$$\overline{Aut(K)} \cap \mathbf{R}_0 \subseteq \overline{Aut(K)} \cap \mathbf{Q} \subseteq \overline{Aut(K)} \cap \mathbf{S} = Aut(K).$$

Proof. The first inclusion comes from Proposition 3.1. The second inclusion follows from the fact that the **Q**-property implies the **S**-property; see section 2.2. The last equality comes from the previous proposition. ∎

This corollary shows that

$$Aut(K) \cap \mathbf{Q} = Aut(K) \cap \mathbf{R}_0 \Rightarrow \overline{Aut(K)} \cap \mathbf{Q} = \overline{Aut(K)} \cap \mathbf{R}_0.$$

This means that, to show the equivalence of $\mathbf{R}_0$- and **Q**-properties in $\overline{Aut(K)}$, it is enough to prove such an equivalence in $Aut(K)$.

Motivated by Murty's result—item (a) in the introduction—we may ask if $\mathbf{R}_0$- and **Q**-properties are equivalent when $L \in \Pi(K)$. While the resolution of this question is our ultimate goal (or a road map), for lack of results describing objects of $\Pi(K)$, in the next two sections we deal with a subset of $\Pi(K)$, namely, $\overline{Aut(K)}$. In particular, we deal with algebra automorphisms and quadratic representations.

4. Algebra automorphisms. Recall that L is an algebra automorphism on V if L is invertible and

$$L(x \circ y) = L(x) \circ L(y)$$

for all x and y. In this section, we describe complementarity properties of such transformations.

To motivate our results, we first consider some examples.

Example 4. Consider $V = R^n$ with the usual inner product and Jordan product (= componentwise product). In this setting, every algebra automorphism of V is given by a permutation matrix. Then Murty's result (see item (a) of the introduction) shows that such an automorphism has the $\mathbf{R}_0$-property if and only if the matrix is the identity matrix; thus **GUS**-, **Q**-, and $\mathbf{R}_0$-properties are equivalent for algebra automorphisms on R^n.

Example 5. Consider $V = \mathcal{S}^n$ with the trace inner product and the usual Jordan product. Then it is known (as a consequence of Schneider's result in [23]) that every algebra automorphism on $\mathcal{S}^n$ is given by

$$L(X) = UXU^T,$$

where U is a (real) orthogonal matrix. In this setting it is known ([2], [11]) that **GUS**-, **Q**-, and $\mathbf{R}_0$-properties are equivalent (to $\pm U$ being positive definite).

Example 6. Consider the Lorentz space $\mathcal{L}^n$. Since the underlying space is R^n, we think of a transformation L on $\mathcal{L}^n$ as given by a matrix

$$L = \begin{bmatrix} a & b^T \\ c & D \end{bmatrix},$$

where $a \in R$, $b, c \in R^{n-1}$, and $D \in R^{(n-1)\times(n-1)}$.

Now suppose that $L \in Aut(\mathcal{L}^n)$. Since L preserves the unit element in $\mathcal{L}^n$, we must have

$$\begin{bmatrix} a & b^T \\ c & D \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

proving $a = 1$ and $c = 0$. As $L \in Aut(\mathcal{L}^n) \subset Aut(\mathcal{L}^n_+)$, by a result of Loewy and Schneider [16], there exists a $\mu > 0$ such that

$$L^T J_n L = \mu\, J_n,$$

where $J_n = \mathrm{diag}(1, -1, -1, \ldots, -1)$. A direct calculation shows that $b = 0$, $\mu = 1$, and $D^T D = I$, and so

$$L = \begin{bmatrix} 1 & 0 \\ 0 & D \end{bmatrix}, \tag{4.1}$$

where D is an orthogonal matrix. (We note that $D = I$ and $D = -I$ are likely candidates.) In this section, we show that for such an automorphism, **GUS**-, **Q**-, and $\mathbf{R}_0$-properties are equivalent.

First we describe the real eigenvalues of an algebra automorphism and a necessary condition for the $\mathbf{R}_0$-property. In what follows, $\sigma(L)$ denotes the spectrum of a linear transformation L.

PROPOSITION 4.1. *Let V be a Euclidean Jordan algebra. If $L \in Aut(V)$, then*

$$\sigma(L) \cap R \subseteq \{-1, 1\}.$$

In particular, $\sigma(L) \cap R = \{1\}$ if and only if $-1 \notin \sigma(L)$.

Proof. As $L(x \circ y) = L(x) \circ L(y)$ for all $x, y \in V$, we have $L(e) = e$, where e is the unit element in V. Thus $1 \in \sigma(L) \cap R$. Now suppose that λ is a real eigenvalue of L (which is nonzero since L is invertible), so that for some nonzero $x \in V$, $L(x) = \lambda x$. It follows that $L(x^2) = \lambda^2 x^2$, and more generally, $L(x^k) = \lambda^k x^k$ for all natural numbers k. Since $x \neq 0 \Rightarrow x^k \neq 0$ (this can be seen by considering the spectral decomposition of x), λ^k is an eigenvalue of L for all k. As V is finite dimensional, $\sigma(L)$ is finite, and so two distinct powers of λ are equal; that is, $\lambda^m = 1$ for some natural number m. As λ is real, we must have $\lambda = \pm 1$. Thus we have $\sigma(L) \cap R \subseteq \{-1, 1\}$. The second statement in the proposition is obvious. □

Remark. The above proposition can also be seen as follows. Given any Euclidean Jordan algebra $(V, \langle \cdot, \cdot \rangle, \circ)$, it is well known that $[x, y] := \text{trace}(x \circ y)$ induces another inner product on V that is compatible with the Jordan product. With respect to this inner product, any algebra automorphism is an orthogonal (i.e., $L^T L = I$). Working with the complexifications of $(V, \langle \cdot, \cdot \rangle)$, $(V, [\cdot, \cdot])$, and L, we see that the spectrum of L (which is independent of the inner product on V) is contained in the unit circle. The above proposition follows immediately from this.

THEOREM 4.2. *Let V be a Euclidean Jordan algebra, $L \in Aut(V)$, and $L \in \mathbf{R}_0$. Then $-1 \notin \sigma(L)$.*

Proof. Suppose $-1 \in \sigma(L)$. Then there exists a vector $0 \neq x \in V$ such that $L(x) = -x$. By Theorem 2.1, there exist unique real numbers $\mu_1, \ldots, \mu_k$, all distinct, and a unique complete system of orthogonal idempotents $f_1, \ldots, f_k$ such that

$$x = \mu_1 f_1 + \cdots + \mu_k f_k.$$

Without loss of generality assume that $\mu_1 \neq 0$. Then

$$-(\mu_1 f_1 + \cdots + \mu_k f_k) = -x = L(x) = \mu_1 L(f_1) + \cdots + \mu_k L(f_k).$$

Because $L \in Aut(V)$, $\{L(f_1), \ldots, L(f_k)\}$ is also a complete system of orthogonal idempotents. Since $\mu_1 \neq -\mu_1$, by the uniqueness property, $-\mu_1 = \mu_i$ and $f_1 = L(f_i)$ for some $1 < i \leq k$. Then f_i is a solution of LCP$(L, K, 0)$, contradicting the $\mathbf{R}_0$-property of L. This completes the proof. □

COROLLARY 4.3. *Let $L \in Aut(V)$, $L \in \mathbf{R}_0$, and $L = L^T$. Then $L = I$.*

Proof. Since $L = L^T$, $\sigma(L) \subseteq R$. As $\{1\} \subseteq \sigma(L) \cap R \subseteq \{-1, 1\}$ and $-1 \notin \sigma(L)$, we have $\sigma(L) = \{1\}$. By the spectral theorem (for operators), $L = I$. □

The following examples show that the converse in the above theorem need not hold.

Example 7. Consider the Euclidean Jordan algebra R^3 with the usual inner product and the Jordan product. Define $L : R^3 \to R^3$ by the matrix

$$L = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

Certainly, $L \in Aut(R^3)$ and $\sigma(L) = \{1, e^{\frac{2\pi i}{3}}, e^{-\frac{2\pi i}{3}}\}$, so $\sigma(L) \cap R = \{1\}$. For $x = (1, 0, 0)^T$, $L(x) = (0, 0, 1)^T$ and $\langle L(x), x \rangle = 0$; thus, $L \notin \mathbf{R}_0$. Note that R^3 is not a simple Jordan algebra.

Example 8. On the simple algebra $\mathcal{S}^3$, let

$$U = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix},$$

and define the transformation L on $\mathcal{S}^3$ by $L(X) = UXU^T$. As U is orthogonal, a simple argument shows that $UX + XU = 0 \Rightarrow X = 0$. Hence -1 is not an eigenvalue of L. By Proposition 4.1, $\sigma(L) \cap R = \{1\}$. As $\pm U$ is not positive definite, a result of Bhimasankaram et al. [2] (see also [11]) shows that L cannot have the $\mathbf{R}_0$-property.

5. A characterization of the global uniqueness property of an algebra automorphism on $\mathcal{L}^n$. In this section, we establish the equivalence of the global uniqueness property and the $\mathbf{R}_0$-property for an automorphism on $\mathcal{L}^n$.

THEOREM 5.1. *For $L \in Aut(\mathcal{L}^n)$, the following are equivalent:*

(i) *L has the* **GUS***-property.*
(ii) *L has the* **P***-property.*
(iii) *L has the* $\mathbf{R}_0$*-property.*
(iv) *L has the* **Q***-property.*
(v) *$-1 \notin \sigma(L)$ (or, equivalently, $-1 \notin \sigma(D)$ with D given in* (4.1)).

Proof. The implications (i) $\Rightarrow$ (ii) $\Rightarrow$ (iii) follow from Theorem 14 in [12] and the definitions. The implication (iii) $\Rightarrow$ (iv) follows from Proposition 3.1. We show that (iv) $\Rightarrow$ (v). Now write L as in (4.1). Suppose L has the **Q**-property and -1 is an eigenvalue of L (as well as that of D), so that for some nonzero $u \in R^{n-1}$

$$Du = -u.$$

Now let x be a solution of $\mathrm{LCP}(L, \mathcal{L}^n_+, q)$, where

$$q = \begin{bmatrix} 0 \\ u \end{bmatrix} \quad \text{and} \quad x = \begin{bmatrix} x_0 \\ \overline{x} \end{bmatrix}.$$

Then

$$x \in \mathcal{L}^n_+,\ L(x) + q \in \mathcal{L}^n_+,\ \text{and}\ \langle x, L(x) + q \rangle = 0,$$

and so

$$x_0 \geq ||\overline{x}||,\ x_0 \geq ||D\overline{x} + u||,\ \text{and}\ x_0^2 + \langle \overline{x}, D\overline{x} + u \rangle = 0.$$

Now, in view of the Cauchy–Schwarz inequality, we have

$$x_0^2 = \langle -\overline{x}, D\overline{x} + u \rangle \leq ||\overline{x}||\,||D\overline{x} + u|| \leq x_0^2.$$

As $q \notin \mathcal{L}^n_+$, x cannot be zero; hence x_0, $\overline{x}$, and $D\overline{x} + u$ are all nonzero. Consequently,

$$D\overline{x} + u = -\theta\overline{x},\ x_0 = ||\overline{x}||,\ \text{and}\ x_0 = ||D\overline{x} + u||$$

for some positive θ. From these, we get $\theta = 1$ and

$$D\overline{x} + u = -\overline{x}.$$

As D is orthogonal, $Du = -u \Rightarrow D^T u = -u$. Thus,

$$-\langle \overline{x}, u \rangle = \langle D\overline{x} + u, u \rangle = \langle D\overline{x}, u \rangle + ||u||^2 = \langle \overline{x}, D^T u \rangle + ||u||^2 = -\langle \overline{x}, u \rangle + ||u||^2.$$

This leads to $||u||^2 = 0$, which is a contradiction. Hence L satisfies (v).

Now for the last implication (v) $\Rightarrow$ (i).

First we show that (v) implies (ii). So assume (v).

Let x be a vector such that x and $L(x)$ operator commute, with $x \circ L(x) \leq 0$. For

$$x = \begin{bmatrix} x_0 \\ \bar{x} \end{bmatrix} \quad \text{and} \quad L(x) = \begin{bmatrix} x_0 \\ D\bar{x} \end{bmatrix}$$

we have

$$x \circ L(x) = \begin{bmatrix} x_0^2 + \langle D\bar{x}, \bar{x} \rangle \\ x_0 D\bar{x} + x_0 \bar{x} \end{bmatrix} \leq 0.$$

Case (i). $D\bar{x} = 0$ or, equivalently, $\bar{x} = 0$ (as D is nonsingular). Then $\begin{bmatrix} x_0^2 \\ 0 \end{bmatrix} \leq 0$, so $x = 0$.

Case (ii). $D\bar{x} \neq 0$. Hence, $\bar{x} \neq 0$ and, by the operator commutativity, $D\bar{x} = \mu\bar{x}$ for some μ. Since $\mu \in \sigma(L) \cap R = \{1\}$, $\mu = 1$; thus, $D\bar{x} = \bar{x}$. Consequently, $x_0^2 + ||\bar{x}||^2 \leq 0$, so $x = 0$.

Hence, L has the **P**-property.

Now to show (i). Take any $q \in \mathcal{L}^n$, and let x and u be two solutions of LCP$(L, \mathcal{L}^n_+, q)$ so that

$$x \geq 0, \quad y = L(x) + q \geq 0, \quad \text{and} \quad \langle x, y \rangle = 0$$

and

$$u \geq 0, \quad v = L(u) + q \geq 0, \quad \text{and} \quad \langle u, v \rangle = 0.$$

We know that x and y operator commute and u and v operator commute (see Proposition 2.2). If we can show that x and v operator commute and u and y operator commute then $x - u$ operator commutes with $y - v$. In this situation,

$$(x - u) \circ L(x - u) = (x - u) \circ (y - v) = -(x \circ v + u \circ y) \leq 0,$$

where the last inequality follows from the fact that if two vectors in K operator commute, then their Jordan product is also in K. At this stage, we can apply the **P**-property (item (ii)) and get $x = u$. Thus (i) is proved provided the following claim can be proved.

Claim. u and y, as well as x and v, operator commute. (Equivalently, $\overline{u}$ and $\overline{y}$ are proportional and $\overline{x}$ and $\overline{v}$ are proportional.) We now proceed to prove the claim. Let

$$x = \begin{bmatrix} x_0 \\ \bar{x} \end{bmatrix}, \; u = \begin{bmatrix} u_0 \\ \bar{u} \end{bmatrix}, \; L(x) + q = \begin{bmatrix} x_0 + q_0 \\ D\bar{x} + \bar{q} \end{bmatrix}, \text{ and } L(u) + q = \begin{bmatrix} u_0 + q_0 \\ D\bar{u} + \bar{q} \end{bmatrix}.$$

We have

$$x_0 \geq ||\bar{x}||, \; u_0 \geq ||\bar{u}||, \; x_0 + q_0 \geq ||D\bar{x} + \bar{q}||, \text{ and } u_0 + q_0 \geq ||D\bar{u} + \bar{q}||,$$

and, moreover,

$$x_0(x_0 + q_0) + \langle D\bar{x} + \bar{q}, \bar{x} \rangle = 0 = u_0(u_0 + q_0) + \langle D\bar{u} + \bar{q}, \bar{u} \rangle.$$

If both x and u are positive, then $L(x) + q = 0 = L(u) + q$ and, by invertibility of L, $x = u$. In this case, the claim is true (because x and y operator commute).

Suppose that (exactly) one of x or u is on the boundary of $\mathcal{L}^n_+$, say, $x > 0$ and $u \in \partial\mathcal{L}^n_+$. Then $x_0 > 0$ and $0 = y = L(x) + q$. If $v > 0$, then $u = 0$, in which case $q > 0$. But then

$$\begin{bmatrix} x_0 \\ D\bar{x} \end{bmatrix} = L(x) = -q < 0$$

implies that $x_0 < 0$, which is a contradiction.

On the other hand, if $v \in \partial\mathcal{L}^n_+$, then

$$u_0 + q_0 = ||D\bar{u} + \bar{q}|| = ||D(\bar{u} - \bar{x})|| = ||\bar{u} - \bar{x}|| \geq ||\bar{u}|| - ||\bar{x}||.$$

As $x_0 = -q_0$, we have $u_0 - x_0 \geq ||\bar{u}|| - ||\bar{x}||$. Thus, $u_0 - ||\bar{u}|| \geq x_0 - ||\bar{x}|| > 0$, so $u > 0$, contradicting our assumption that $u \in \partial\mathcal{L}^n_+$.

Hence we may assume that both x and u are on the boundary of $\mathcal{L}^n_+$. Similarly, by considering L^{-1}, we may assume that both y and v are on the boundary of $\mathcal{L}^n_+$. (Note that $L^{-1} \in Aut(\mathcal{L}^n)$, y, and v are solutions of $\mathrm{LCP}(L^{-1}, K, L^{-1}q)$.)

So at this stage, $x, u, y, v \in \partial\mathcal{L}^n_+$. Thus,

$$x_0 = ||\bar{x}||,\ x_0 + q_0 = ||D\bar{x} + \bar{q}||,\ u_0 = ||\bar{u}||, \quad \text{and} \quad u_0 + q_0 = ||D\bar{u} + \bar{q}||.$$

Case 1. $\bar{x} = 0$. Then $x = 0$ and $q_0 = ||\bar{q}||$.

Subcase 1.1. $\bar{u} = 0$. Then $u = 0$, so $x = u = 0$, and the claim holds.

Subcase 1.2. $\bar{u} \neq 0$. Since u and v operator commute,

$$\bar{v} = D\bar{u} + \bar{q} = \beta\bar{u} \quad \text{for some} \quad \beta \in R. \tag{5.1}$$

Also, $u_0 + q_0 = ||D\bar{u} + \bar{q}|| \leq ||\bar{u}|| + q_0 \leq u_0 + q_0$. The equality in the triangle inequality gives, along with $\bar{u} \neq 0$,

$$D\bar{u} = \theta\bar{q} \quad \text{for some} \quad \theta > 0. \tag{5.2}$$

Now, by (5.2),

$$v = L(u) + q = \begin{bmatrix} u_0 + q_0 \\ D\bar{u} + \bar{q} \end{bmatrix} = \begin{bmatrix} u_0 + ||q|| \\ \theta\bar{q} + \bar{q} \end{bmatrix}$$

and

$$y = L(x) + q = 0 + q = \begin{bmatrix} q_0 \\ \bar{q} \end{bmatrix}.$$

Combining (5.1) and (5.2), we get $\theta\bar{q} + \bar{q} = \beta\bar{u}$. Hence, $\bar{q} = \frac{1}{\theta+1}\beta\bar{u}$ $(\theta > 0)$, so $\bar{q}$ and $\bar{u}$ are proportional, and u and q operator commute. Thus, u and $y\,(= q)$ operator commute. Also, $x = 0$ and v operator commute. Therefore, in Case 1, the claim holds.

Case 2. $\bar{x} \neq 0$.

Subcase 2.1. $\bar{u} = 0$. This is analogous to Subcase 1.2.

Subcase 2.2. $\bar{u} \neq 0$. Now we have $x, u, y, v \in \partial\mathcal{L}^n_+$, and $\bar{x} \neq 0$, $\bar{u} \neq 0$. Again, the complementarity property, hence operator commutativity, gives

$$D\bar{x} + \bar{q} = \alpha\bar{x} \quad \text{and} \quad D\bar{u} + \bar{q} = \gamma\bar{u}$$

for some α, $\gamma \in R$. Also,

$$x_0 + q_0 = ||D\bar{x} + \bar{q}|| = |\alpha| \cdot ||\bar{x}|| \quad \text{and} \quad u_0 + q_0 = ||D\bar{u} + \bar{q}|| = |\gamma| \cdot ||\bar{u}||.$$

Since $y = L(x) + q = \begin{bmatrix} x_0 + q_0 \\ \alpha\bar{x} \end{bmatrix}$ and $0 = \langle x, y \rangle = x_0(x_0 + q_0) + \alpha||\bar{x}||^2$, we have $\alpha \leq 0$. Similarly, $\gamma \leq 0$. Since $x_0 + q_0 = |\alpha| \cdot ||\bar{x}||$ and $x \in \partial\mathcal{L}^n_+$, $q_0 = |\alpha| \cdot ||\bar{x}|| - x_0 = (|\alpha| - 1)||\bar{x}||$; likewise, $q_0 = (|\gamma| - 1)||\bar{u}||$. Also, $\bar{q} = \alpha\bar{x} - D\bar{x} = \gamma\bar{u} - D\bar{u}$. Then

$$D(\bar{x} - \bar{u}) = \alpha\bar{x} - \gamma\bar{u}.$$

Since D is orthogonal, $||\bar{x} - \bar{u}|| = ||\alpha\bar{x} - \gamma\bar{u}||$ and $q_0 = (|\alpha| - 1)||\bar{x}|| = (|\gamma| - 1)||\bar{u}||$, $\alpha, \gamma \leq 0$. Put $\lambda := -\alpha$ and $\mu := -\gamma$, so $\lambda, \mu \geq 0$,

$$D(\overline{x} - \overline{u}) = \mu\overline{u} - \lambda\overline{x}, \ ||\bar{x} - \bar{u}|| = ||\mu\bar{u} - \lambda\bar{x}||, \quad \text{and} \quad (\lambda - 1)||\bar{x}|| = (\mu - 1)||\bar{u}||.$$

If $\overline{x}$ and $\overline{u}$ are proportional, then so are $\overline{y}\,(= \alpha\overline{x})$ and $\overline{u}$, and $\overline{x}$ and $\overline{v}\,(= \gamma\overline{u})$. In this setting, the claim holds. Suppose $\overline{x}$ and $\overline{u}$ are not proportional. Then, as the signs of $\lambda - 1$ and $\mu - 1$ are the same, two-dimensional (Euclidean) geometric considerations show that the quadrilateral with vertices $\bar{x}$, $\lambda\bar{x}$, $\mu\bar{u}$, and $\bar{u}$ (which is part of a triangle) can be a parallelogram only when $\lambda = 1$ and $\mu = 1$ (see the appendix for an algebraic proof). In this situation,

$$D(\bar{x} - \bar{u}) = -(\bar{x} - \bar{u}),$$

violating the condition (v) that $-1 \notin \sigma(D)$. Thus we have the claim in Subcase 2.2. This shows that the claim is proved for Case 2. Hence (v) $\Rightarrow$ (i). □

Remark. One may wonder if the above result (Theorem 5.1) is true for algebra automorphisms on a direct sum of $\mathcal{L}^n$'s. To address this, let $V := \mathcal{L}^{n_1} \oplus \mathcal{L}^{n_2} \oplus \cdots \oplus \mathcal{L}^{n_k}$ and $L \in Aut(V)$. If L is "diagonal," that is, if $L = L_1 \oplus L_2 \oplus \cdots \oplus L_k$, where $L_i : \mathcal{L}^{n_i} \to \mathcal{L}^{n_i}$, then $L_i \in Aut(\mathcal{L}^{n_i})$ for all i. In this situation, Theorem 5.1 extends to L on V. On the other hand, if L is not "diagonal," Theorem 5.1 may not extend to L. This can be seen by modifying Example 7:

Put $V = \mathcal{L}^n \oplus \mathcal{L}^n \oplus \mathcal{L}^n$ and $L(x, y, z) = (y, z, x)$. It is easily seen that $L \in Aut(V)$ and $-1 \notin \sigma(L)$, yet $L \notin \mathbf{R}_0$ (because $(e, 0, 0)$ is a solution of LCP$(L, K, 0)$).

6. Quadratic representations. The algebra automorphisms of a Euclidean Jordan algebra, studied in the previous section, form an important subclass of $Aut(K)$. In this section, we consider another important subclass of $\overline{Aut(K)}$, namely, quadratic representations. Given any element a in the Euclidean Jordan algebra V, the quadratic representation of a is defined by

$$P_a(x) := 2a \circ (a \circ x) - a^2 \circ x.$$

It is known that P_a is a self-adjoint linear transformation on V and belongs to $Aut(K)$ when a is invertible. Our main result below establishes the equivalence of global uniqueness, global solvability, and the $\mathbf{R}_0$-properties for such transformations. Our motivation comes from the following.

Consider $V = \mathcal{S}^n$ and $K = \mathcal{S}^n_+$. Then a linear transformation L on V belongs to $Aut(\mathcal{S}^n_+)$ if and only if there exists an invertible matrix $A \in R^{n \times n}$ such that

$$L(X) = AXA^T \quad (X \in \mathcal{S}^n);$$

see [23]. For such transformations, it has been shown in [2] (see also [11]) that global uniqueness and $\mathbf{R}_0$-properties are equivalent and that these properties hold if and only if $\pm A$ is positive definite.

Now when $A \in \mathcal{S}^n$, the above transformation coincides with P_A given by:

$$P_A(X) = AXA.$$

In [21], Sampangi Raman considers P_A on $\mathcal{S}^n$ and shows that the global solvability and the $\mathbf{R}_0$-properties are equivalent and that these properties hold if and only if $\pm A$ is positive definite. Working on $\mathcal{L}^n$, Malik and Mohan [17] prove a similar result for quadratic representations on $\mathcal{L}^n$.

In our main result below, we extend these two results to arbitrary Euclidean Jordan algebras. We remark that the crucial idea of our analysis comes from [21].

We recall the following from [6] (see Props. II.3.1, II.3.2, and III.2.2).

PROPOSITION 6.1. *Let $a \in V$. Then the following statements hold:*

(i) *a is invertible if and only if P_a is invertible.*
(ii) *$P_{P_a(x)} = P_a P_x P_a$.*
(iii) *If a is invertible, then $P_a(K) = K$. Hence $P_a(K) \subseteq K$ for all $a \in V$.*
(iv) *If for some x, $P_a(x)$ is invertible, then a is invertible.*

The following lemmas are needed to prove our main theorem. These lemmas and their proofs are somewhat similar, except for technical details, to those in [21], [17].

LEMMA 6.2. *Let V be any Euclidean Jordan algebra. Let a be invertible in V with spectral decomposition given by*

$$a = a_1 e_1 + a_2 e_2 + \cdots + a_r e_r.$$

Define $|a| := |a_1|e_1 + |a_2|e_2 + \cdots + |a_r|e_r$ and $s = \varepsilon_1 e_1 + \varepsilon_2 e_2 + \cdots + \varepsilon_r e_r$, where $\varepsilon_i = sign\,(a_i)$. If P_a has the $\mathbf{Q}$-property, then so does P_s.

Proof. Let $b := \sqrt{|a|}$. Then $P_b(s) = a$ (by using the definition) and $P_a = P_b P_s P_b$ using item (ii) in the previous lemma. Assume that P_a has the $\mathbf{Q}$-property, and let $q \in V$. Let x be a solution of LCP(P_a, K, r), where $r = P_b(q)$. Then $x \geq 0$, $y := P_a(x) + r \geq 0$, and $\langle x, y\rangle = 0$. From $P_a = P_{P_b(s)} = P_b P_s P_b$, we have $P_b^{-1} y = P_s(P_b(x)) + q$. Using item (iii) in the above lemma and the self-adjointness of P_b, we have $u := P_b(x) \geq 0$, $v := P_b^{-1} y \geq 0$, and $\langle u, v\rangle = \langle P_b(x), P_b^{-1} y\rangle = \langle x, y\rangle = 0$. This means that LCP$(P_s, K, q)$ has a solution, proving the result. □

LEMMA 6.3. *Let $\{e_1, e_2, \ldots, e_r\}$ be a Jordan frame in V and $x = \sum_1^r x_i e_i + \sum_{i<j} x_{ij}$ be the Peirce decomposition of an element $x \in V$ with respect to this Jordan frame. Let*

$$s = e_1 + e_2 + \cdots + e_k - (e_{k+1} + \cdots + e_r)$$

for some k, with $1 \leq k < r$ and $0 \neq q_{k\,k+1} \in V_{k\,k+1}$. Then the following hold:

(a) *$P_s(x) = \sum_1^r x_i e_i + \sum_\beta x_{ij} - \sum_\alpha x_{ij}$, where $\alpha := \{(i,j) : 1 \leq i \leq k, k+1 \leq j \leq r\}$ and $\beta := \{(i,j) : 1 \leq i < j \leq r\} \setminus \alpha$.*
(b) *The $(k\,k+1)$-term in the Peirce decomposition of $x \circ P_s(x)$ is zero.*
(c) *The $(k\,k{+}1)$-term in the Peirce decomposition of $x \circ q_{k\,k+1}$ is $\frac{1}{2}(x_k + x_{k+1}) q_{k\,k+1}$.*
(d) *If $x \geq 0$, then $x_k e_k + x_{k+1} e_{k+1} + x_{k\,k+1} \geq 0$.*

Proof. Let $f = e_1 + e_2 + \cdots + e_k$ so that $s = 2f - e$, where e is the unit element in V. Then $P_s(x) = 2s \circ (s \circ x) - s^2 \circ x$. Since $s^2 = e$ and $s = 2f - e$, simplification leads to $P_s(x) = 8f \circ (f \circ x) - 8f \circ x + x$. Using the properties

$$e_l \circ e_i = \delta_{il} e_l \quad \text{and} \quad e_l \circ x_{ij} = \frac{1}{2} x_{ij} \text{ if } l \in \{i, j\}, \text{ or } 0 \text{ if } l \notin \{i, j\},$$

we get

$$\begin{aligned} f \circ x &= x_1 e_1 + \frac{1}{2}(x_{12} + x_{13} + \cdots + x_{1r}) \\ &+ x_2 e_2 + \frac{1}{2}(x_{12} + x_{23} + x_{24} + \cdots + x_{2r}) \\ &+ x_3 e_3 + \frac{1}{2}(x_{13} + x_{23} + x_{34} + \cdots + x_{3r}) \\ &+ \cdots + \\ &+ x_k e_k + \frac{1}{2}(x_{1k} + x_{2k} + \cdots + x_{k-1\,k} + x_{k\,k+1} \cdots + x_{kr}). \end{aligned}$$

We note that, in the above expression, the term x_{ij} for $1 \le i \le k$ and $k+1 \le j \le r$ appears only once and the term x_{ij} for $1 \le i < j \le k$ appears twice. Hence

$$f \circ x = \sum_1^k x_i e_i + \sum_{1 \le i < j \le k} x_{ij} + \frac{1}{2} \sum_{1 \le i \le k, k+1 \le j \le r} x_{ij}.$$

From this, we get $f \circ (f \circ x) = \sum_1^k x_i e_i + \sum_{1 \le i < j \le k} x_{ij} + \frac{1}{4} \sum_\alpha x_{ij}$. Using $P_s(x) = 8f \circ (f \circ x) - 8f \circ x + x$, a simple calculation leads to item (a).

We now prove item (b). Consider any element y with its Peirce decomposition:

$$y = \sum_1^r y_i e_i + \sum_{m<n} y_{mn}.$$

Then, in view of the properties of the spaces V_{ij}, the $(k\,k+1)$-term in the Peirce decomposition of $x \circ y$ is obtained by adding all terms of the form $x_{ij} \circ y_{mn}$ and $x_{mn} \circ y_{ij}$, where

$$k \in \{i, j\},\ k+1 \in \{m, n\}, \text{ and } |\{i, j\} \cap \{m, n\}| = 1.$$

This sum reduces to

$$\begin{aligned} &\sum_{1 \le i < k} x_{ik} \circ y_{i\,k+1} + \sum_{1 \le i < k} y_{ik} \circ x_{i\,k+1} \\ &+ x_{kk} \circ y_{k\,k+1} + y_{kk} \circ x_{k\,k+1} \\ &+ x_{k\,k+1} \circ y_{k+1\,k+1} + y_{k\,k+1} \circ x_{k+1\,k+1} \\ &+ \sum_{k+1 < i \le r} x_{ki} \circ y_{k+1\,i} + \sum_{k+1 < i \le r} y_{ki} \circ x_{k+1\,i}. \end{aligned}$$

Now, when $y = P_s(x)$, we have $y_{ij} = -x_{ij}$ for $(i, j) \in \alpha$ and $y_{ij} = x_{ij}$ for $(i, j) \notin \alpha$. Putting these in the above sum and simplifying, we get item (b).

Upon putting $y = q_{k\,k+1}$, we see that the $(k\,k+1)$-term in the Peirce decomposition of $x \circ y$ is

$$x_{kk} \circ q_{k\,k+1} + x_{k+1\,k+1} \circ q_{k\,k+1} = \frac{1}{2}(x_k + x_{k+1}) q_{k\,k+1},$$

which is item (c).

Now we prove item (d). Suppose $x \ge 0$. Let

$$V_{\{e_k, e_{k+1}\}} = \{x \in V : x \circ (e_k + e_{k+1}) = x\}.$$

It is known (see Prop. IV.1.1 in [6]) that this is a Euclidean Jordan algebra and its corresponding symmetric cone is given by (Thm. 3.1, [13])

$$V^+_{\{e_k,e_{k+1}\}} := \{x \in K : x \circ (e_k + e_{k+1}) = x\}.$$

Let y belong to this (sub)cone. Then its Peirce decomposition in V with respect to $\{e_1, e_2, \ldots, e_r\}$ is given by (see Lem. 20, [12])

$$y = y_k e_k + y_{k+1} e_{k+1} + y_{k\,k+1}.$$

Then, using the orthogonality properties of spaces V_{ij}, we have

$$0 \le \langle x, y\rangle = \langle x_k e_k + x_{k+1} e_{k+1} + x_{k\,k+1}, y\rangle.$$

As y is arbitrary, we see from the self-duality of $V^+_{\{e_k,e_{k+1}\}}$ that $x_k e_k + x_{k+1} e_{k+1} + x_{k\,k+1} \in V^+_{\{e_k,e_{k+1}\}}$ and, in particular, belongs to K. □

PROPOSITION 6.4. *Suppose V is simple. Let $\{e_1, e_2, \ldots, e_r\}$ be a Jordan frame in V. Let $s = e_1 + e_2 + \cdots + e_k - (e_{k+1} + \cdots + e_r)$ for some k, with $1 \le k < r$. Then P_s does not have the* **Q***-property.*

Proof. As V is simple, $V_{k\,k+1}$ is nontrivial (see Prop. IV.2.3, [6]). Let $0 \ne q_{k\,k+1} \in V_{k\,k+1}$. We claim that $\mathrm{LCP}(P_s, K, q_{k\,k+1})$ has no solution. If possible, let x be a solution of this problem. Then $x \ge 0$, $y := P_s(x) + q_{k\,k+1} \ge 0$, and $x \circ y = 0$. Applying the previous lemma to this setting, we get

$$0 = (x \circ y)_{k\,k+1} = (x \circ P_s(x))_{k\,k+1} + (x \circ q_{k\,k+1})_{k\,k+1} = \frac{1}{2}(x_k + x_{k+1}) q_{k\,k+1}.$$

As $q_{k\,k+1} \ne 0$, we must have $x_k + x_{k+1} = 0$. But $x \ge 0$ implies that $x_k e_k + x_{k+1} e_{k+1} + x_{k\,k+1} \ge 0$. So x_k and x_{k+1} are both nonnegative, and hence $x_k = x_{k+1} = 0$. By Prop. 3.2 in [13], $x_{k\,k+1} = 0$. Now $y \ge 0$ implies that $y_k e_k + y_{k+1} e_{k+1} + y_{k\,k+1} \ge 0$. From $y = P_s(x) + q_{k\,k+1}$ and the above lemma, we get $x_k e_k + x_{k+1} e_{k+1} - x_{k\,k+1} + q_{k\,k+1} \ge 0$. As $0 = x_k = x_{k+1}$ and $x_{k\,k+1} = 0$, we have (by Prop. 3.2 in [13]) $q_{k\,k+1} = 0$, which is a contradiction. Hence $\mathrm{LCP}(P_s, K, q_{k\,k+1})$ has no solution. □

THEOREM 6.5. *Let V be any Euclidean Jordan algebra and $a \in V$. Then the following are equivalent:*

(1) *P_a is positive definite on V.*
(2) *P_a has the* **GUS***-property.*
(3) *P_a has the* **P***-property.*
(4) *P_a has the* $\mathbf{R}_0$*-property.*
(5) *P_a has the* **Q***-property.*

If, in addition, V is simple, then the above conditions are further equivalent to

(6) *$\pm a \in K^\circ$.*

Proof. The implications (1) ⇒ (2) ⇒ (3) ⇒ (4) hold for any linear transformation on V; see [12]. Now suppose that (4) holds. Since $P_a(K) \subseteq K$, by Prop. 3.1, P_a has the **Q**-property. Hence (5) holds. Now suppose that (5) holds. If x is a solution of $\mathrm{LCP}(P_a, K, -e)$, then $P_a(x) - e \ge 0$, and hence $P_a(x) > 0$. By Item (iv) in Proposition 6.1, a is invertible. Let $a = a_1 e_1 + a_2 e_2 + \cdots + a_r e_r$ be the spectral decomposition of a. Note that each a_i is nonzero. By Lemma 6.2, P_s has the **Q**-property where $s = \varepsilon_1 e_1 + \varepsilon_2 e_2 + \cdots + \varepsilon_r e_r$, with $s_i = sign\,(a_i)$. First suppose that V is simple. Then by Proposition 6.4, $s_i = 1$ for all i or $s_i = -1$ for all i. This means that $\pm a \in K^\circ$. Since $P_a = P_{-a}$, we may assume that $a > 0$. Then $P_a = P_{P_{\sqrt{a}}e} = P_{\sqrt{a}} P_e P_{\sqrt{a}} = P^2_{\sqrt{a}}$,

and so P_a is positive semidefinite on V. As P_a is invertible and symmetric (recall that a is invertible), P_a must be positive definite on V. Thus in the case of simple V, conditions (1)–(6) are equivalent. When V is not simple, we show that (5) is equivalent to (1) by decomposing V as a product of simple Euclidean Jordan algebras (cf. Theorem 2.4). Write $V = V^{(1)} \times V^{(2)} \times \cdots \times V^{(m)}$, where each $V^{(k)}$ is simple. Let $a = (a^{(1)}, a^{(2)}, \ldots, a^{(m)})$ in this product. As $P_a = P_{a^{(1)}} \times P_{a^{(2)}} \times \cdots \times P_{a^{(m)}}$, we see that each $P_{a^{(k)}}$ has the **Q**-property in $V^{(k)}$. By what has been proved, $\pm a^{(k)} > 0$ in $V^{(k)}$ and $P_{a^{(k)}}$ is positive definite on $V^{(k)}$. It follows that P_a is positive definite on V. This completes the proof. □

7. Lyapunov-like transformations. A real square matrix A is said to be a **Z**-matrix if all its off-diagonal entries are nonpositive. If $A \in R^{n\times n}$, then this property is equivalent to

$$x, y \in R^n_+,\ \langle x, y\rangle = 0 \Rightarrow \langle Ax, y\rangle \leq 0.$$

This concept can be extended to symmetric cones: Following [14], we say a linear transformation $L : V \to V$ has the **Z**-property if

$$x, y \in K,\ \langle x, y\rangle = 0 \Rightarrow \langle L(x), y\rangle \leq 0.$$

It has been shown (see [25], [5]) that this property is equivalent to $e^{-tL} \in \Pi(K)$ for all $t \geq 0$.

The recent article [14] contains properties of such transformations; in particular, it is shown in that paper that L has the global solvability property (item (β) of the introduction) if and only if L is positive stable. Examples of **Z**-transformations include both Lyapunov and Stein transformations on $\mathcal{S}^n$. We now say that a linear transformation L on V is a *Lyapunov-like* transformation if both L and $-L$ have the **Z**-property; that is,

$$x, y \in K,\ \langle x, y\rangle = 0 \Rightarrow \langle L(x), y\rangle = 0.$$

Recently, Damm [4] has shown that for $\mathcal{S}^n$ and $\mathcal{H}^n$ (the space of all $n \times n$ Hermitian matrices over complex numbers) L has the above property if and only if it is a Lyapunov transformation (that is, it is of the form L_A for some square matrix A). While the form of a Lyapunov-like transformation on a general Euclidean Jordan algebra is not known, it can be easily shown [26] that, on $\mathcal{L}^n$, a matrix is Lyapunov-like if and only if it is of the form

$$\begin{bmatrix} a & b^T \\ b & D \end{bmatrix},$$

where $a \in R$, $D \in R^{(n-1)\times(n-1)}$, with $D + D^T = 2aI$. For any Euclidean Jordan algebra V and $a \in V$, the transformation L_a (called a Lyapunov transformation) defined by

$$L_a(x) = a \circ x$$

is also a Lyapunov-like transformation.

Extending the result in item (1) of the introduction, we present the following global uniqueness result.

THEOREM 7.1. *Let $L : V \to V$ be a Lyapunov-like transformation. Then L has the* **GUS***-property if and only if L is positive stable (that is, all its eigenvalues have positive real parts) and positive semidefinite.*

Proof. Suppose L has the **GUS**-property. Then it has the global solvability property and so by Thm. 7 in [14], L is positive stable. Also, by Thm. 4.1 in [27],

$$\langle L(c), c\rangle \geq 0$$

for any primitive idempotent c in V. Now for any $x \in V$, we have the spectral decomposition

$$x = \sum_1^r \lambda_i e_i,$$

where $\{e_1, e_2, \ldots, e_r\}$ is a Jordan frame. Since L is a Lyapunov-like transformation, we have $\langle L(e_i), e_j\rangle = 0$ for all $i \neq j$, and so

$$\langle L(x), x\rangle = \sum_{i,j} \lambda_i \lambda_j \langle L(e_i), e_j\rangle = \sum_i \lambda_i^2 \langle L(e_i), e_i\rangle \geq 0.$$

This proves that L is positive semidefinite.

Now assume that L is positive stable and positive semidefinite. Because of positive stability, for every q, LCP(L, K, q) has a solution; see Thms. 6 and 7 in [14]. We now prove uniqueness. Fix q, and suppose that x and u are two solutions of LCP(L, K, q) so that

$$x \geq 0,\ y := L(x) + q \geq 0,\ \text{and } \langle x, y\rangle = 0$$

and

$$u \geq 0,\ v := L(u) + q \geq 0,\ \text{and } \langle u, v\rangle = 0.$$

Now, as L is positive semidefinite, the solution set of LCP(L, K, q) is convex (Thm. 2.3.5, [7]). So for any $t \in [0, 1]$, $tx + (1-t)u$ is also a solution of LCP(L, K, q). Writing out the complementarity conditions, we see that

$$\langle x, v\rangle = 0 = \langle u, y\rangle.$$

We conclude (by Proposition 2.2) that x and u operator commute with both y and v, and $x \circ v = 0 = u \circ y$; hence $z := x - u$ operator commutes with $y - v = L(z)$, and $z \circ L(z) = 0$. In this situation, there exists a Jordan frame $\{e_1, e_2, \ldots, e_r\}$ and scalars μ_i such that

$$z = \mu_1 e_1 + \mu_2 e_2 + \cdots + \mu_l e_l \quad \text{and} \quad L(z) = \mu_{l+1} e_{l+1} + \mu_{l+2} e_{l+2} + \cdots + \mu_r e_r$$

for some l between 1 and r. We then have

$$L(z) = \mu_1 L(e_1) + \mu_2 L(e_2) + \cdots + \mu_l L(e_l) = \mu_{l+1} e_{l+1} + \mu_{l+2} e_{l+2} + \cdots + \mu_r e_r. \tag{7.1}$$

Now for any i between $l+1$ and r, and k between 1 and l, we have $\langle e_k, e_i\rangle = 0$ and $\langle L(e_k), e_i\rangle = 0$. From (7.1), we get

$$\mu_i ||e_i||^2 = 0.$$

This implies that $L(z) = 0$. Since L is positive stable, it is invertible, and so $z = 0$, thus proving the the uniqueness of solution for LCP(L, K, q). Hence L has the **GUS**-property. □

Concluding remarks. In this article, we have proved that global uniqueness, global solvability, and the $\mathbf{R}_0$-properties are equivalent for algebra automorphisms over $\mathcal{L}^n$ and for quadratic representations over any Euclidean Jordan algebra. We have given a characterization of the global uniqueness property for Lyapunov-like transformations. All of the transformations considered in this paper are related to symmetric-cone-invariant transformations. Motivated by our results, we pose the following problems:

(1) Do global uniqueness and $\mathbf{R}_0$-properties coincide for algebra automorphisms on general Euclidean Jordan algebras?
(2) Do global solvability (i.e., the $\mathbf{Q}$-property) and $\mathbf{R}_0$-properties coincide for a cone automorphism? For an element of $\Pi(K)$?

8. Appendix. Here we justify an assertion made in the proof of Theorem 5.1.

LEMMA 8.1. *Let $\bar{x}$ and $\bar{u}$ be two nonzero vectors in a real inner product space that are not proportional. Assume that, for some nonnegative scalars λ and μ,*

$$||\bar{x} - \bar{u}|| = ||\mu\bar{u} - \lambda\bar{x}|| \quad \text{and} \quad (\lambda - 1)||\bar{x}|| = (\mu - 1)||\bar{u}||.$$

Then $\lambda = \mu = 1$.

Proof. The second equality shows that $(\lambda - 1)(\mu - 1) \geq 0$. Assume now that $(\lambda - 1)(\mu - 1) > 0$.

Case 1. $\lambda > 1$ and $\mu > 1$. Expanding the first equality above, applying the Cauchy–Schwarz inequality, and using the second equality, we get

$$\begin{aligned} 0 &= (\mu^2 - 1)||\bar{u}||^2 + (\lambda^2 - 1)||\bar{x}||^2 + 2(\lambda\mu - 1)\langle -\bar{x}, \bar{u}\rangle \\ &\geq (\mu^2 - 1)||\bar{u}||^2 + (\lambda^2 - 1)||\bar{x}||^2 - 2(\lambda\mu - 1)||\bar{u}|| \cdot ||\bar{x}|| \\ &= (\mu^2 - 1)||\bar{u}||^2 + (\lambda^2 - 1)\frac{(\mu - 1)^2}{(\lambda - 1)^2}||\bar{u}||^2 - 2\frac{(\lambda\mu - 1)}{(\lambda - 1)}(\mu - 1)||\bar{u}||^2. \end{aligned} \tag{8.1}$$

Simplifying (8.1), we get

$$0 \geq (\mu^2 - 1)(\lambda - 1)^2 + (\lambda^2 - 1)(\mu - 1)^2 - 2(\lambda\mu - 1)(\mu - 1)(\lambda - 1) = 0.$$

Since $\lambda\mu > 1$, we have equality in the Cauchy–Schwarz inequality $\langle \bar{x}, \bar{u}\rangle \leq ||\bar{x}|| \cdot ||\bar{u}||$. This means that the vectors $\bar{x}$ and $\bar{u}$ are proportional, which is a contradiction.

Case 2. $0 \leq \lambda < 1$ and $0 \leq \mu < 1$.

We omit the proof as it is similar to Case 1.

Contradictions obtained in both cases imply that $\lambda = \mu = 1$. ∎

REFERENCES

[1] A. BERMAN AND R.J. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, SIAM, Philadelphia, 1994.
[2] P. BHIMASANKARAM, A.L.N. MURTHY, G.S.R. MURTHY, AND T. PARTHASARATHY, *Complementarity Problems and Positive Definite Matrices*, Research report, Indian Statistical Institute, Street No. 8, Habshiguda, Hyderabad 500 007, India, 2000 (revised June 27, 2001).
[3] R.W. COTTLE, J.-S. PANG, AND R.E. STONE, *The Linear Complementarity Problem*, Academic, Boston, 1992.
[4] T. DAMM, *Positive groups on H^n are completely positive*, Linear Algebra Appl., 393 (2004), pp. 127–137.
[5] L. ELSNER, *Quasimonotonie and ungleischungen in halbgeordneten Räumen*, Linear Algebra Appl., 8 (1974), pp. 249–261.

[6] J. Faraut and A. Korányi, *Analysis on Symmetric Cones*, Oxford University Press, Oxford, 1994.
[7] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Springer-Verlag, New York, 2003.
[8] M.S. Gowda, *An analysis of zero set and global error bound properties of a piecewise affine function via its recession function*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 594–609.
[9] M.S. Gowda and T. Parthasarathy, *Complementarity forms of theorems of Lyapunov and Stein, and related results*, Linear Algebra Appl., 320 (2000), pp. 131–144.
[10] M.S. Gowda and Y. Song, *On semidefinite linear complementarity problems*, Math. Program., 88 (2000), pp. 575–587.
[11] M.S. Gowda, Y. Song, and G. Ravindran, *On some interconnections between strict monotonicity, GUS, and P properties in semidefinite linear complementarity problems*, Linear Algebra Appl., 370 (2003), pp. 355–368.
[12] M.S. Gowda, R. Sznajder, and J. Tao, *Some P-properties for linear transformations on Euclidean Jordan algebras*, Special issue on Positivity, Linear Algebra Appl., 393 (2004), pp. 203–232.
[13] M.S. Gowda and R. Sznajder, *Automorphism invariance of* $\mathbf{P}$ *and* **GUS** *properties of linear transformations in Euclidean Jordan algebras*, Math. Oper. Res., 31 (2006), pp. 109–123.
[14] M.S. Gowda and J. Tao, *Z-transformations on proper and symmetric cones*, Math. Program., to appear.
[15] S. Karamardian, *An existence theorem for the complementarity problem*, J. Optim. Theory Appl., 19 (1976), pp. 227–232.
[16] R. Loewy and H. Schneider, *Positive operators on the n-dimensional ice-cream cone*, J. Math. Anal. Appl., 49 (1975), pp. 375–392.
[17] M. Malik and S.R. Mohan, *On* $\mathbf{Q}$ *and* $\mathbf{R}_0$ *properties of a quadratic representation in linear complementarity problems over the second-order cone*, Linear Algebra Appl., 379 (2005), pp. 85–97.
[18] K.G. Murty, *On the number of solutions to the linear complementarity problem and spanning properties of complementary cones*, Linear Algebra Appl., 5 (1972), pp. 65–108.
[19] T. Parthasarathy, D. Sampangi Raman, and B. Sriparna, *Relationship between strong monotonicity,* $\mathbf{P}_2$*-property and the GUS property in semidefinite LCPs*, Math. Oper. Res., 27 (2002), pp. 326–331.
[20] S.M. Robinson, *Normal maps induced by linear transformations*, Math. Oper. Res., 17 (1992), pp. 691–714.
[21] D. Sampangi Raman, *Some Contributions to Semidefinite Linear Complementarity Problem*, Ph.D. thesis, Indian Statistical Institute, Chennai, India, 2003.
[22] S.H. Schmieta and F. Alizadeh, *Extension of primal-dual interior point algorithms to symmetric cones*, Math. Program., 96 (2003), pp. 409–438.
[23] H. Schneider, *Positive operators and an inertia theorem*, Numer. Math., 7 (1965), pp. 11–17.
[24] S. Scholtes, *Introduction to Piecewise Differentiable Equations*, Preprint 53 1994, Institute für Statistik und Mathematische Wirtschaftstheorie, Universität Karlsruhe, 7500 Karlsruhe, Germany, 1994.
[25] H. Schneider and M. Vidyasagar, *Cross-positive matrices*, SIAM J. Numer. Anal., 7 (1970), pp. 508–519.
[26] J. Tao, *private communication*, 2006.
[27] J. Tao and M.S. Gowda, *Some P-properties for nonlinear transformations on Euclidean Jordan algebras*, Math. Oper. Res., 30 (2005), pp. 985–1004.

SIAM J. OPTIM.
Vol. 18, No. 2, pp. 482–506

STOCHASTIC R_0 MATRIX LINEAR COMPLEMENTARITY PROBLEMS*

HAITAO FANG†, XIAOJUN CHEN‡, AND MASAO FUKUSHIMA§

Abstract. We consider the expected residual minimization formulation of the stochastic R_0 matrix linear complementarity problem. We show that the involved matrix being a stochastic R_0 matrix is a necessary and sufficient condition for the solution set of the expected residual minimization problem to be nonempty and bounded. Moreover, local and global error bounds are given for the stochastic R_0 matrix linear complementarity problem. A stochastic approximation method with acceleration by averaging is applied to solve the expected residual minimization problem. Numerical examples and applications of traffic equilibrium and system control are given.

Key words. stochastic linear complementarity problem, R_0 matrix, expected residual minimization

AMS subject classifications. 90C33, 90C15

DOI. 10.1137/050630805

1. Introduction.

Let $(\Omega, \mathcal{F}, P)$ be a probability space, where Ω is a subset of $\mathbb{R}^m$, and $\mathcal{F}$ is a σ-algebra generated by $\{\Omega \cap U : U \text{ is an open set in } \mathbb{R}^m\}$. We consider the stochastic linear complementarity problem (SLCP):

$$x \geq 0,\ M(\omega)x + q(\omega) \geq 0,\quad x^T(M(\omega)x + q(\omega)) = 0,$$

where $M(\omega) \in \mathbb{R}^{n\times n}$ and $q(\omega) \in \mathbb{R}^n$ for $\omega \in \Omega$. We denote this problem by SLCP$(M(\omega), q(\omega))$ for short. Throughout this paper, we assume that $M(\omega)$ and $q(\omega)$ are measurable functions of ω with the following property:

$$E\{\|M(\omega)^T M(\omega)\|\} < \infty \quad \text{and} \quad E\{\|q(\omega)\|^2\} < \infty,$$

where E stands for the expectation. If Ω only contains a single realization, then the SLCP reduces to the standard LCP. For the standard LCP, much effort has been made in developing theoretical analysis for the existence of a solution, numerical methods for finding a solution, and applications in engineering and economics [5,7,9]. On the other hand, in many practical applications, some data in the LCP cannot be known with certainty. The SLCP is aimed at a practical treatment of the LCP under uncertainty. However, only a little attention has been paid to the SLCP in the literature.

In general, there is no x satisfying the SLCP$(M(\omega), q(\omega))$ for almost all $\omega \in \Omega$. A deterministic formulation for the SLCP provides a decision vector which is optimal

*Received by the editors May 6, 2005; accepted for publication (in revised form) December 10, 2006; published electronically May 29, 2007. This work is partly supported by the Information Research Center for Development of Knowledge Society Infrastructure, Graduate School of Informatics, Kyoto University, Japan.

http://www.siam.org/journals/siopt/18-2/63080.html

†Key Laboratory of Systems and Control, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, 100080, China (htfang@iss.ac.cn). The work of this author is also supported by National Natural Science Foundation of China.

‡Department of Mathematical Sciences, Hirosaki University, Hirosaki 036-8561, Japan (chen@cc.hirosaki-u.ac.jp). The work of this author is also supported by a Grant-in-Aid from Japan Society for the Promotion of Science.

§Department of Applied Mathematics and Physics, Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan (fuku@i.kyoto-u.ac.jp). The work of this author is also supported by a Grant-in-Aid from Japan Society for the Promotion of Science.

in a certain sense. Different deterministic formulations may yield different solutions that are optimal in different senses.

Gürkan, Özge, and Robinson [12] considered the sample-path approach for stochastic variational inequalities and provided convergence theory and applications for the approach. When applied to the SLCP$(M(\omega), q(\omega))$, the approach is the same as the *expected value* (EV) method, which uses the expected function of the random function $M(\omega)x + q(\omega)$ and solves the deterministic problem

$$x \geq 0, \quad E\{M(\omega)x + q(\omega)\} \geq 0,\ x^T E\{M(\omega)x + q(\omega)\} = 0.$$

Using a simulation-based algorithm in [12], we can find a solution of this problem.

Recently, Chen and Fukushima [3] proposed a new deterministic formulation called the *expected residual minimization* (ERM) method, which is to find a vector $x \in \mathbb{R}^n_+$ that minimizes the expected residual of the SLCP$(M(\omega), q(\omega))$, i.e.,

$$\min_{x \in \mathbb{R}^n_+} E\{\|\Phi(x, \omega)\|^2\}, \tag{1.1}$$

where $\Phi : \mathbb{R}^n \times \Omega \to \mathbb{R}^n$ is defined by

$$\Phi(x, \omega) = \begin{pmatrix} \phi([M(\omega)x]_1 + q_1(\omega), x_1) \\ \vdots \\ \phi([M(\omega)x]_n + q_n(\omega), x_n) \end{pmatrix},$$

and $[x]_i$ denotes the ith component of the vector x. Here $\phi : \mathbb{R}^2 \to \mathbb{R}$ is an NCP function which has the property

$$\phi(a, b) = 0 \iff a \geq 0,\ b \geq 0,\ ab = 0.$$

Various NCP functions have been studied for solving complementarity problems [7]. In this paper, we will concentrate on the "min" function

$$\phi(a, b) = \min(a, b).$$

Similar results can be obtained for other NCP functions, such as the Fischer–Burmeister (FB) function [10], which have the same growth behavior as the "min" function.

Let ERM$(M(\cdot), q(\cdot))$ denote problem (1.1) and define

$$G(x) = \int_\Omega \|\Phi(x, \omega)\|^2 dF(\omega), \tag{1.2}$$

where $F(\omega)$ is the distribution function of ω. Then ERM$(M(\cdot), q(\cdot))$ is rewritten as

$$\min\ G(x) \qquad \text{s.t.}\ \ x \geq 0. \tag{1.3}$$

Recall that an $n \times n$ matrix A is called an R_0 matrix if

$$x \geq 0,\ Ax \geq 0,\ x^T Ax = 0 \qquad \Longrightarrow \qquad x = 0.$$

It is known [5, Theorem 3.9.23] that the solution set of the standard LCP(A, b)

$$x \geq 0,\ Ax + b \geq 0,\ x^T(Ax + b) = 0$$

is bounded for every $b \in \mathbb{R}^n$, if and only if A is an R_0 matrix. In addition, when A is a P_0 matrix, the LCP(A, b) has a nonempty solution set if and only if A is an R_0 matrix [5, Theorem 3.9.22]. Example 1 in [3] shows that the solution set of LCP$(M(\bar{\omega}), q(\bar{\omega}))$ being nonempty and bounded for some $\bar{\omega} \in \Omega$ does not imply that the ERM$(M(\cdot), q(\cdot))$ has a solution. The following results on the existence of a solution of ERM$(M(\cdot), q(\cdot))$ are given in [3].

(i) If $M(\cdot)$ is continuous in ω and there is an $\bar{\omega} \in \Omega$ such that $M(\bar{\omega})$ is an R_0 matrix, then the solution set of ERM$(M(\cdot), q(\cdot))$ is nonempty and bounded.

(ii) When $M(\omega) \equiv M$, the solution set of ERM$(M(\cdot), q(\cdot))$ is nonempty and bounded for any $q(\cdot)$ if and only if M is an R_0 matrix.

In this paper, we substantially extend and refine the results established in [3]. In particular, we introduce the concept of a stochastic R_0 matrix and show that $M(\cdot)$ being a stochastic R_0 matrix is a necessary and sufficient condition for the solution set of ERM$(M(\cdot), q(\cdot))$ to be nonempty and bounded. Moreover, we will extend the local and global error bound results for the R_0 matrix LCP given by Mangasarian and Ren [16] to the stochastic R_0 matrix LCP in the ERM formulation.

Throughout the paper, the norm $\|\cdot\|$ denotes the Euclidean norm and $\mathbb{R}^n_+ = \{x \in \mathbb{R}^n : x \geq 0\}$. For a given vector $x \in \mathbb{R}^n$, we denote $I(x) = \{i : x_i = 0\}$ and $J(x) = \{i : x_i \neq 0\}$. For vectors $x, y \in \mathbb{R}^n$, $\min(x, y)$ denotes the vector with components $\min(x_i, y_i)$, $i = 1, \ldots, n$.

The remainder of the paper is organized as follows: In section 2, the definition and some properties of a stochastic R_0 matrix are given. In section 3, we show a necessary and sufficient condition for the existence of a minimizer of the ERM problem with an arbitrary $q(\cdot)$ is that $M(\cdot)$ is a stochastic R_0 matrix. In section 4, the differentiability of G is considered. Some optimality conditions and error bounds of the ERM problem are given in section 5. In section 6, we use a stochastic approximation method [2, 14] with acceleration by averaging [18] to solve the general ERM problem, and use a Newton-type method to solve the ERM problem with $M(\omega) \equiv M$. Furthermore, applications to traffic equilibrium and system control are provided. Preliminary numerical results show that the ERM formulation has various advantages.

2. Stochastic R_0 matrix. A stochastic R_0 matrix is formally defined as follows.

DEFINITION 2.1. *$M(\cdot)$ is called a stochastic R_0 matrix if*

$$x \geq 0,\ M(\omega)x \geq 0,\ x^T M(\omega)x = 0, \quad a.e. \implies x = 0.$$

If Ω only contains a single realization, then the definition of a stochastic R_0 matrix reduces to that of an R_0 matrix.

Let G be defined by (1.2). We call $x^* \in \mathbb{R}^n_+$ a *local solution* of the ERM$(M(\cdot),\ q(\cdot))$, if there is $\gamma > 0$ such that $G(x) \geq G(x^*)$ for all $x \in \mathbb{R}^n_+ \cap B(x^*, \gamma) := \{x : \|x - x^*\| \leq \gamma\}$, and call x^* a *global solution* of ERM$(M(\cdot),\ q(\cdot))$, if $G(x) \geq G(x^*)$ for all $x \in \mathbb{R}^n_+$.

THEOREM 2.2. *The following statements are equivalent.*

(i) *$M(\cdot)$ is a stochastic R_0 matrix.*

(ii) *For any $x \geq 0$ $(x \neq 0)$, at least one of the following two conditions is satisfied:*
   (a) *$P\{\omega : [M(\omega)x]_i \neq 0\ \} > 0$ for some $i \in J(x)$;*
   (b) *$P\{\omega : [M(\omega)x]_i < 0\} > 0$ for some $i \in I(x)$.*

(iii) *ERM$(M(\cdot), q(\cdot))$ with $q(\omega) \equiv 0$ has zero as its unique global solution.*

Proof. The proof is given in the order (i) $\Rightarrow$ (iii) $\Rightarrow$ (ii) $\Rightarrow$ (i).

(i) $\Rightarrow$ (iii): It is easy to see that zero is a global solution of ERM$(M(\cdot), q(\cdot))$ with $q(\omega) \equiv 0$, since $G(x) \geq 0$ for all $x \in \mathbb{R}^n_+$ and $G(0) = 0$. Now we show the uniqueness of

the solution. Let $\bar{x} \in \mathbb{R}^n_+$ be an arbitrary vector such that $G(\bar{x}) = 0$. By the definition of G, we have

$$\Phi(\bar{x}, \omega) = \min(M(\omega)\bar{x}, \bar{x}) = 0, \quad \text{a.e.},$$

which implies

$$\bar{x} \geq 0, \quad M(\omega)\bar{x} \geq 0, \quad \bar{x}^T M(\omega)\bar{x} = 0, \quad \text{a.e.}$$

By the definition of a stochastic R_0 matrix, we deduce $\bar{x} = 0$.

(iii) $\Rightarrow$ (ii): Suppose (ii) does not hold, that is, there exists a nonzero $x^0 \geq 0$ such that

$$P\{\omega : \ [M(\omega)x^0]_i = 0\} = 1 \text{ for all } i \in J(x^0),$$

$$P\{\omega : \ [M(\omega)x^0]_i \geq 0\} = 1 \text{ for all } i \in I(x^0).$$

Then it follows from $q(\omega) \equiv 0$ that $G(x^0) = 0$. Moreover, it is easy to see that for any $\lambda > 0$, λx^0 is a solution of ERM$(M(\cdot), 0)$, i.e., zero is not the unique solution of ERM$(M(\cdot), 0)$.

(ii) $\Rightarrow$ (i): Assume that there exists $x \neq 0$ such that $x \geq 0$, $M(\omega)x \geq 0$, and $x^T M(\omega)x = 0$, a.e. Then, since $x^T M(\omega)x = 0$, we have for almost all ω, $[M(\omega)x]_i = 0$ for all $i \in J(x)$ and $[M(\omega)x]_i \geq 0$ for all $i \in I(x)$. This contradicts (ii). ∎

For $\nu > 0$, let us denote $B_\Omega(\bar{\omega}, \nu) := \{\omega : \|\omega - \bar{\omega}\| < \nu\}$ and

$$\text{supp}\Omega := \left\{ \bar{\omega} \in \Omega : \ \int_{B_\Omega(\bar{\omega}, \nu) \cap \Omega} dF(\omega) > 0 \text{ for any } \nu > 0 \right\}.$$

Here suppΩ is called the support set of Ω. When Ω consists of countable discrete points, i.e., $\Omega = \{\omega_1, \ldots, \omega_i, \ldots\}$ and $P(\omega_i) = p_i > 0$ for all i, we have supp$\Omega = \Omega$. In the case that there is a density function ρ such that $dF(\omega) = \rho(\omega)d\omega$, we have supp$\Omega = \bar{S}$, where $\bar{S}$ is the closure of set $S = \{\omega \in \Omega : \ \rho(\omega) > 0\}$.

COROLLARY 2.3. *Suppose that $M(\omega)$ is a continuous function of ω. Then $M(\cdot)$ is a stochastic R_0 matrix if and only if for any $x \geq 0\,(x \neq 0)$, at least one of the following two conditions is satisfied:*

(a) *there exists $\bar{\omega} \in$ suppΩ such that $[M(\bar{\omega})x]_i \neq 0$ for some $i \in J(x)$;*

(b) *there exists $\bar{\omega} \in$ suppΩ such that $[M(\bar{\omega})x]_i < 0$ for some $i \in I(x)$.*

Proof. By the continuity of $M(\omega)$ and the definition of suppΩ, conditions (a) and (b) in this corollary imply (a) and (b) in Theorem 2.2 (ii), respectively. ∎

COROLLARY 2.4. *Suppose that $M(\omega)$ is a continuous function of ω and $M(\bar{\omega})$ is an R_0 matrix for some $\bar{\omega} \in$ suppΩ. Then $M(\cdot)$ is a stochastic R_0 matrix.*

The following example shows that the condition that $M(\cdot)$ is a stochastic R_0 matrix is weaker than the condition that $M(\omega)$ is continuous in ω and there is an $\bar{\omega} \in$ suppΩ such that $M(\bar{\omega})$ is an R_0 matrix.

Example 2.1. Let

$$M(\omega) = \begin{pmatrix} -2\omega & \omega - |\omega| & 0 \\ 0 & \omega + |\omega| & -2\omega \\ 0 & 0 & 0 \end{pmatrix},$$

where $\omega \in \Omega = [-0.5, 0.5]$ and ω is uniformly distributed on Ω. Clearly, for $\omega < 0$, $M(\omega) = \begin{pmatrix} -2\omega & 2\omega & 0 \\ 0 & 0 & -2\omega \\ 0 & 0 & 0 \end{pmatrix}$. Then $x = (1, 1, 0)^T$ satisfies $M(\omega)x = 0$. On the other hand,

for $\omega > 0$, $M(\omega) = \begin{pmatrix} -2\omega & 0 & 0 \\ 0 & 2\omega & -2\omega \\ 0 & 0 & 0 \end{pmatrix}$. Then $x = (0,1,1)^T$ satisfies $M(\omega)x = 0$. In this example, there is no $\omega \in \Omega$ such that $M(\omega)$ is an R_0 matrix. However, $M(\cdot)$ is a stochastic R_0 matrix as verified by Theorem 2.2 (ii). For any $x \geq 0$ with $x \neq 0$, if $x_1 \neq 0$, then for any $\omega > 0$, $[M(\omega)x]_1 = -2\omega x_1 < 0$. If $x_1 = 0$ but $x_2 \neq 0$, then for any $\omega < 0$, $[M(\omega)x]_1 = 2\omega x_2 < 0$. If only $x_3 \neq 0$, then for any $\omega > 0$, $[M(\omega)x]_2 = -2\omega x_3 < 0$.

The following proposition shows a relation between $M(\cdot)$ and $\bar{M} := E\{M(\omega)\}$.

PROPOSITION 2.5. *If $\bar{M}$ is an R_0 matrix, then $M(\cdot)$ is a stochastic R_0 matrix.*

Proof. If $M(\cdot)$ were not a stochastic R_0 matrix, then by Theorem 2.2 (ii), there exists $x \geq 0$ such that $x \neq 0$ and, for almost all ω, $[M(\omega)x]_i = 0$ for $i \in J(x)$ and $[M(\omega)x]_i \geq 0$ for $i \in I(x)$. Therefore, $[\bar{M}x]_i = 0$ for $i \in J(x)$ and $[\bar{M}x]_i \geq 0$ for $i \in I(x)$. This is impossible, since $\bar{M}$ is an R_0 matrix. ∎

This proposition implies that for any given $\tilde{M}$, if $\tilde{M}$ is an R_0 matrix, then $M(\cdot) = \tilde{M} + M_0(\cdot)$ with $E\{M_0(\omega)\} = 0$ is a stochastic R_0 matrix. The converse of this proposition is not true. The next proposition gives a way to construct a stochastic R_0 matrix $M(\cdot)$ from a given $\tilde{M}$ which is not necessarily an R_0 matrix. Let

$$\Xi(M) := \{x : x \geq 0, x \neq 0,\ [Mx]_i = 0, i \in J(x) \text{ and } [Mx]_i \geq 0, i \in I(x)\}. \tag{2.1}$$

Obviously, if $\Xi(\tilde{M}) = \emptyset$, then $\tilde{M}$ is an R_0 matrix, and hence, by Proposition 2.5, $M(\cdot) = \tilde{M} + M_0(\cdot)$ with $E\{M_0(\omega)\} = 0$ is a stochastic R_0 matrix.

PROPOSITION 2.6. *Let $\tilde{M}$ and $M_0(\cdot)$ be such that $\Xi(\tilde{M}) \neq \emptyset$ and $E\{M_0(\omega)\} = 0$. Suppose that for any $x \in \Xi(\tilde{M})$, at least one of the following two conditions is satisfied:*

(1) *For some $i \in J(x)$, $E\{([M_0(\omega)x]_i)^2\} > 0$;*

(2) *For some $i \in I(x)$, $P\{\omega : [M_0(\omega)x]_i < -b\} > 0$ for any $b > 0$.*

Then $M(\cdot) = \tilde{M} + M_0(\cdot)$ is a stochastic R_0 matrix.

Proof. For $x \in \Xi(\tilde{M})$, these two conditions imply that the conditions in Theorem 2.2 (ii) hold for $M(\cdot)$. For $x \notin \Xi(\tilde{M})$, the same conditions also hold trivially. So $M(\cdot)$ is a stochastic R_0 matrix. ∎

This proposition suggests a way to obtain a stochastic R_0 matrix $M(\cdot)$ from an arbitrary matrix $\tilde{M}$. Specifically, we can construct a simple stochastic matrix $M_0(\cdot)$ such that $\tilde{M} + M_0(\cdot)$ is a stochastic R_0 matrix, as illustrated in the following example.

Example 2.2. We consider the following matrix [3]:

$$\tilde{M} = \begin{pmatrix} 0 & 0 & 1 & -2 & -3 \\ 0 & 0 & 1 & -6 & -3 \\ -1 & -1 & 0 & 0 & 0 \\ 2 & 6 & 0 & 0 & 0 \\ 3 & 3 & 0 & 0 & 0 \end{pmatrix},$$

which arises from a linear programming problem in [13]. Clearly, $\tilde{M}$ is not an R_0 matrix, and $\Xi(\tilde{M}) = \{x : x = (0,0,\lambda,\alpha,\beta)^T, \lambda > 0, \alpha, \beta \geq 0, \lambda - 6\alpha - 3\beta \geq 0\}$. Let ω_0 be a random variable whose distribution is $\mathcal{N}(0,1)$. Let

$$M_0(\omega_0) = \begin{pmatrix} 0 & 0 & 0.5\omega_0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ -0.5\omega_0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Then for any $b > 0$, $P\{\omega_0 : [M_0(\omega_0)x]_1 < -b\} > 0$ holds for any $x \in \Xi(\tilde{M})$. Hence, by Proposition 2.6, $\tilde{M} + M_0(\cdot)$ is a stochastic R_0 matrix.

The following proposition shows that the sum of a stochastic R_0 matrix $M(\cdot)$ and a matrix $M_1(\cdot)$ with $E\{M_1(\omega_1)\} = 0$ yields a stochastic R_0 matrix.

PROPOSITION 2.7. *Let $\omega = (\omega_0, \omega_1)$ and $\hat{M}(\omega) = M(\omega_0) + M_1(\omega_1)$, where $M(\cdot)$ is a stochastic R_0 matrix, $E\{M_1(\omega_1)\} = 0$, and $M(\omega_0)$ is independent of $M_1(\omega_1)$. Then $\hat{M}(\cdot)$ is a stochastic R_0 matrix.*

Proof. If $\tilde{M} := E\{M(\omega_0)\}$ is an R_0 matrix, then from $E\{M_1(\omega_1)\} = 0$ and Proposition 2.5, $M(\cdot)+M_1(\cdot)$ is a stochastic R_0 matrix. Otherwise, let $M_0(\omega_0) = M(\omega_0)-\tilde{M}$ and choose any $x \in \Xi(\tilde{M})$. Suppose that the first condition of Proposition 2.6 holds for $M_0(\omega_0)$. Since $M(\omega_0)$ is independent of $M_1(\omega_1)$, we have

$$E\{([(M_0(\omega_0) + M_1(\omega_1))x]_i)^2\} = E\{([M_0(\omega_0)x]_i)^2\} + E\{([M_1(\omega_1)x]_i)^2\} > 0$$

for some $i \in J(x)$. Now, suppose that the second condition of Proposition 2.6 holds for $M_0(\omega_0)$, i.e., $P\{\omega_0 : [M_0(\omega_0)x]_i < -b\} > 0$ for some $i \in I(x)$. Note that

$$\begin{aligned} &P\{\omega : [(M_0(\omega_0) + M_1(\omega_1))x]_i < -b\} \\ &\quad \geq P\{(\omega_0, \omega_1) : [M_0(\omega_0)x]_i < -b \text{ and } [M_1(\omega_1)x]_i \leq 0\} \\ &\quad = P\{\omega_0 : [M_0(\omega_0)x]_i < -b\} P\{\omega_1 : [M_1(\omega_1)x]_i \leq 0\}. \end{aligned}$$

Since $E\{[M_1(\omega_1)x]_i\} = 0$, we have $P\{\omega_1 : [M_1(\omega_1)x]_i \leq 0\} > 0$. Thus, we have

$$P\{\omega : [(M_0(\omega_0) + M_1(\omega_1))x]_i < -b\} > 0,$$

i.e., the second condition of Proposition 2.6 also holds for $M_0(\omega_0) + M_1(\omega_1)$. Since

$$\hat{M}(\omega) = M(\omega_0) + M_1(\omega_1) = \tilde{M} + M_0(\omega_0) + M_1(\omega_1),$$

Proposition 2.6 ensures that $\hat{M}(\cdot)$ is a stochastic R_0 matrix. ∎

3. Boundedness of solution set. In this section, the boundedness of the solution set of the ERM problem (1.3) is studied.

THEOREM 3.1. *Let $q(\cdot)$ be arbitrary. Then $G(x) \to \infty$ as $\|x\| \to \infty$ with $x \in \mathbb{R}^n_+$ if and only if $M(\cdot)$ is a stochastic R_0 matrix.*

Proof. First, we prove the "if" part. For simplicity, we denote $|x| = (|x_1|, \ldots, |x_n|)^T$ and $\text{sign}(x) = (\text{sign}(x_1), \ldots, \text{sign}\ (x_n))^T$ for a vector x, where

$$\text{sign}(x_i) = \begin{cases} 1, & x_i > 0, \\ 0, & x_i = 0, \\ -1, & x_i < 0. \end{cases}$$

Note that for any $a, b \in \mathbb{R}$, we have

$$\begin{aligned} 2\min(a, b) &= a + b - \text{sign}(a - b)(a - b) \\ &= (1 - \text{sign}(a - b))a + (1 + \text{sign}(a - b))b \end{aligned}$$

and

$$\begin{aligned} 4(\min(a, b))^2 &= a(1 - \text{sign}(a - b))^2 a + b(1 + \text{sign}(a - b))^2 b + 2b(1 - \text{sign}^2(a - b))a \\ &= 2a(1 - \text{sign}(a - b))a + 2b(1 + \text{sign}(a - b))b. \end{aligned}$$

For any $x \in \mathbb{R}^n$ and $\omega \in \Omega$, we define the diagonal matrix

$$D(x,\omega) = \mathrm{diag}(\mathrm{sign}(M(\omega)x + q(\omega) - x)).$$

Then we have

$$\|\Phi(x,\omega)\|^2 = \frac{1}{2}[(M(\omega)x + q(\omega))^T(I - D(x,\omega))(M(\omega)x + q(\omega)) + x^T(I + D(x,\omega))x]. \tag{3.1}$$

Consider an arbitrary $x \geq 0$ with $\|x\| = 1$. Suppose condition (a) in Theorem 2.2 (ii) holds. Choose $i \in J(x)$ such that $P\{\omega : [M(\omega)x]_i \neq 0\} > 0$. Then there exists a sufficiently large $K > 0$ such that $P\{\omega : [M(\omega)x]_i \neq 0,\ |q_i(\omega)| \leq K\} > 0$.

First, consider the case where $P\{\omega : [M(\omega)x]_i < x_i,\ |q_i(\omega)| \leq K\} > 0$. Let

$$\Omega_1 := \{\omega : [M(\omega)x]_i < (1-\delta)x_i,\ |q_i(\omega)| \leq K\},$$

where $\delta > 0$. Then we have $P\{\Omega_1\} > 0$ whenever δ is sufficiently small. Moreover, for any sufficiently large $\lambda > 0$, $\mathrm{sign}(\lambda[M(\omega)x]_i + q_i(\omega) - \lambda x_i) = -1$ for any $\omega \in \Omega_1$. Therefore, by (1.2) and (3.1), we have

$$G(\lambda x) \geq \int_{\Omega_1} (\lambda[M(\omega)x]_i + q_i(\omega))^2 dF(\omega) \to \infty \text{ as } \lambda \to \infty. \tag{3.2}$$

Next, consider the case where $P\{\omega : [M(\omega)x]_i > x_i, |q_i(\omega)| \leq K\} > 0$. Let

$$\Omega_2 := \{\omega : [M(\omega)x]_i > (1+\delta)x_i, |q_i(\omega)| \leq K\}.$$

Then we have $P(\Omega_2) > 0$ for a sufficiently small $\delta > 0$. Moreover, for any sufficiently large $\lambda > 0$, $\mathrm{sign}(\lambda[M(\omega)x]_i + q_i(\omega) - \lambda x_i) = 1$ for $\omega \in \Omega_2$. Hence we have

$$G(\lambda x) \geq \int_{\Omega_2} (\lambda x_i)^2 dF(\omega) \to \infty \text{ as } \lambda \to \infty. \tag{3.3}$$

Finally, consider the case where $P\{\omega : [M(\omega)x]_i = x_i, |q_i(\omega)| \leq K\} > 0$. Let

$$\Omega_3 := \{\omega : [M(\omega)x]_i = x_i, |q_i(\omega)| \leq K\}.$$

Then we have

$$G(\lambda x) \geq \int_{\Omega_3} \{(\lambda x_i + q_i(\omega))^2 1_{\{q_i(\omega)<0\}} + (\lambda x_i)^2 1_{\{q_i(\omega)\geq 0\}}\} dF(\omega) \to \infty \text{ as } \lambda \to \infty. \tag{3.4}$$

Combining (3.2)–(3.4), we see that $G(\lambda x) \to \infty$ as $\lambda \to \infty$.

Now, suppose condition (b) in Theorem 2.2 (ii) holds. Choose $i \in I(x)$ such that $P\{\omega : [M(\omega)x]_i < 0\} > 0$. Let

$$\Omega_4 := \{\omega :\ [M(\omega)x]_i < -\delta,\ |q_i(\omega)| < K\}.$$

Then we have $P\{\Omega_4\} > 0$ for any sufficiently small $\delta > 0$ and sufficiently large $K > 0$. Moreover, for any $\lambda > 0$ large enough, $\lambda[M(\omega)x]_i + q_i(\omega) < 0$ for $\omega \in \Omega_4$. Thus we have

$$(1 - \mathrm{sign}(\lambda[M(\omega)x]_i + q_i(\omega)))(\lambda[M(\omega)x]_i + q_i(\omega))^2 = 2(\lambda[M(\omega)x]_i + q_i(\omega))^2,$$

which yields

$$G(\lambda x) \geq \int_{\Omega_4} (\lambda[M(\omega)x]_i + q_i(\omega))^2 dF(\omega) \to \infty \text{ as } \lambda \to \infty.$$

Since x is an arbitrary nonzero vector such that $x \geq 0$, we deduce from the above arguments that $G(x) \to \infty$ as $\|x\| \to \infty$ with $x \geq 0$, provided the statement (ii) in Theorem 2.2 holds.

Let us turn to proving the "only if" part. Suppose that $M(\cdot)$ is not a stochastic R_0 matrix, i.e., there exists $x \geq 0$ with $x \neq 0$ such that $[M(\omega)x]_i = 0$ for all $i \in J(x)$ and $[M(\omega)x]_i \geq 0$ for all $i \in I(x)$, a.e. For any $\lambda > 0$, from (1.2) and (3.1), we have

$$G(\lambda x) = \frac{1}{2} \sum_{i=1}^{n} E\{(1 - \text{ sign}(\lambda[M(\omega)x]_i + q_i(\omega) - \lambda x_i))(\lambda[M(\omega)x]_i + q_i(\omega))^2$$

$$+ (1 + \text{sign}(\lambda[M(\omega)x]_i + q_i(\omega) - \lambda x_i))(\lambda x_i)^2\}. \tag{3.5}$$

The ith term of the right-hand side of (3.5) with $x_i \neq 0$ equals

$$E\{(1 - \text{sign}(q_i(\omega) - \lambda x_i))q_i(\omega)^2 + (1 + \text{sign}(q_i(\omega) - \lambda x_i))(\lambda x_i)^2\}$$

$$= 2E\{q_i(\omega)^2 1_{\{q_i(\omega\} \leq \lambda x_i\}} + (\lambda x_i)^2 1_{\{q_i(\omega) > \lambda x_i\}}\} \leq 2E\{q_i(\omega)^2\},$$

while the ith term of the right-hand side of (3.5) with $x_i = 0$ equals

$$E\{(1 - \text{sign}(\lambda[M(\omega)x]_i + q_i(\omega)))(\lambda[M(\omega)x]_i + q_i(\omega))^2\}$$

$$= 2E\{(\lambda[M(\omega)x]_i + q_i(\omega))^2 1_{\{\lambda[M(\omega)x]_i < -q_i(\omega)\}}\} \leq 2E\{q_i(\omega)^2\},$$

where the last inequality follows from $0 > \lambda[M(\omega)x]_i + q_i(\omega) \geq q_i(\omega)$, implying $(\lambda[M(\omega)x]_i + q_i(\omega))^2 \leq q_i(\omega)^2$. So, we obtain

$$G(\lambda x) \leq E\{\|q(\omega)\|^2\} \text{ for any } \lambda > 0.$$

Since $x \geq 0$ with $x \neq 0$, this particularly implies that G is bounded above on a nonnegative ray in $\mathbb{R}^n_+$. This completes the proof of the "only if" part. □

The solution set of ERM$(M(\cdot), q(\cdot))$ may be bounded even if $M(\cdot)$ is not a stochastic R_0 matrix. It depends on the distribution of $q(\omega)$, as shown in the following two propositions.

PROPOSITION 3.2. *If $M(\cdot)$ is not a stochastic R_0 matrix, $P\{\omega : q_i(\omega) > 0\} > 0$ for some $i \in J(x)$, and $P\{\omega : q_i(\omega) \geq 0\} = 1$ for all $i \in I(x)$, where $x \neq 0$ is any nonnegative vector at which the conditions* (a) *and* (b) *in Theorem* 2.2 (ii) *fail to hold, then the solution set of ERM$(M(\cdot), q(\cdot))$ is bounded.*

Proof. Note that

$$G(0) = E\{\|\Phi(0, \omega)\|^2\} = \sum_{i=1}^{n} E\{q_i(\omega)^2 1_{\{q_i(\omega) < 0\}}\}. \tag{3.6}$$

For any nonnegative vector $x \neq 0$ satisfying the conditions (a) and (b) in Theorem 2.2 (ii), the proof of Theorem 3.1 indicates that

$$G(\lambda x) \to \infty \text{ as } \lambda \to 0. \tag{3.7}$$

Let $x \neq 0$ be any nonnegative vector which does not satisfy the conditions (a) and (b) in Theorem 2.2 (ii), i.e., $[M(\omega)x]_i = 0$ for $i \in J(x)$, and $[M(\omega)x]_i \geq 0$ for $i \in I(x)$, a.e. Then by (3.1), we have

$$G(\lambda x) = \sum_{i\in J(x)} E\{[(1-\operatorname{sign}(q_i(\omega)-\lambda x_i))q_i(\omega)^2 + (1+\operatorname{sign}(q_i(\omega)-\lambda x_i))(\lambda x_i)^2]/2\}$$

$$= \sum_{i\in J(x)} \{E\{q_i(\omega)^2\} - E\{1_{\{q_i(\omega)-\lambda x_i>0\}}[q_i(\omega)^2 - (\lambda x_i)^2]\}\}, \tag{3.8}$$

where the first equality follows from the assumption that $P\{\omega : q_i(\omega) \geq 0\} = 1$ for $i \in I(x)$ and hence $[M(\omega)x]_i + q_i(\omega) \geq 0$, a.e., for $i \in I(x)$. Note that

$$0 \leq E\{1_{\{q_i(\omega)-\lambda x_i>0\}}[q_i(\omega)^2 - (\lambda x_i)^2]\} = E\{1_{\{q_i(\omega)>\lambda x_i\}}[q_i(\omega)^2 - (\lambda x_i)^2]\}$$

$$\leq E\{1_{\{q_i(\omega)>\lambda x_i\}} q_i(\omega)^2\} \to 0 \text{ as } \lambda \to \infty,$$

which together with (3.8) implies

$$\lim_{\lambda\to\infty} G(\lambda x) = \sum_{i\in J(x)} E\{q_i(\omega)^2\}. \tag{3.9}$$

On the other hand, for any nonzero $x \geq 0$, we have

$$\sum_{i=1}^{n} E\{q_i(\omega)^2 1_{\{q_i(\omega)<0\}}\} = \sum_{i\in J(x)} E\{q_i(\omega)^2 1_{\{q_i(\omega)<0\}}\} < \sum_{i\in J(x)} E\{q_i(\omega)^2\}, \tag{3.10}$$

where the equality follows from the assumption that $P\{\omega : q_i(\omega) \geq 0\} = 1$ for all $i \in I(x)$, and the strict inequality follows from the assumption that $P\{\omega : q_i(\omega) > 0\} > 0$ for some $i \in J(x)$. Combining (3.6), (3.9), and (3.10), we have

$$G(0) < \lim_{\lambda\to+\infty} G(\lambda x). \tag{3.11}$$

Let $\Lambda := \{x \in \mathbb{R}^n_+ : G(x) \leq G(0)\}$. From (3.7) and (3.11), we have $\sup_{x\in\Lambda} \|x\| < +\infty$. Since any solution belongs to Λ, this implies that the solution set is bounded. ∎

PROPOSITION 3.3. *If $M(\cdot)$ is not a stochastic R_0 matrix and, for any i, $P\{\omega : -b \leq q_i(\omega) < 0\} = 1$ for some $b > 0$ and $P\{\omega : q_i(\omega) \neq 0$ and $M(\omega)x^0]_i = 0\} = 0$, where $x^0 \neq 0$ is any nonnegative vector at which the conditions* (a) *and* (b) *in Theorem* 2.2 (ii) *fail to hold, then the solution set of ERM$(M(\cdot), q(\cdot))$ is empty or unbounded.*

Proof. Let $x^0 \neq 0$ be any nonnegative vector which does not satisfy the conditions (a) and (b) in Theorem 2.2 (ii). From (1.2) and (3.1), we have

$$G(\lambda x^0) = \sum_{i=1}^{n} E\{[(1-\operatorname{sign}(\lambda[M(\omega)x^0]_i + q_i(\omega) - \lambda x_i^0))(\lambda[M(\omega)x^0]_i + q_i(\omega))^2$$

$$+ (1+\operatorname{sign}(\lambda[M(\omega)x^0]_i + q_i(\omega) - \lambda x_i))(\lambda x_i^0)^2]/2\}. \tag{3.12}$$

For every $i \in J(x^0)$, we have $[M(\omega)x^0]_i = 0$ and $q_i(\omega) = 0$, a.e., and hence the ith term of the right-hand side of (3.12) is zero for any $\lambda > 0$. For every $i \in I(x^0)$, we have $[M(\omega)x^0]_i \geq 0$ and $q_i(\omega) < 0$, a.e., which implies

$$\begin{aligned} &E\{(1 - \operatorname{sign}(\lambda[M(\omega)x^0]_i + q_i(\omega)))(\lambda[M(\omega)x^0]_i + q_i(\omega))^2\} \\ &\quad = 2E\{(\lambda[M(\omega)x^0]_i + q_i(\omega))^2 1_{\{\lambda[M(\omega)x^0]_i < -q_i(\omega), [M(\omega)x^0]_i > 0\}}\} \\ &\quad\quad + 2E\{q_i^2(\omega) 1_{\{[M(\omega)x^0]_i = 0\}}\}. \end{aligned} \tag{3.13}$$

By assumption, the second term on the right-hand side of (3.13) is zero for any $\lambda > 0$, and

$$\begin{aligned} &E\{(\lambda[M(\omega)x^0]_i + q_i(\omega))^2 1_{\{\lambda[M(\omega)x^0]_i < -q_i(\omega), [M(\omega)x^0]_i > 0\}}\} \\ &\quad \leq b^2 P\{\omega : 0 < \lambda[M(\omega)x^0]_i < b\} \to 0 \text{ as } \lambda \to \infty. \end{aligned}$$

Therefore, we obtain

$$\lim_{\lambda \to +\infty} G(\lambda x^0) = 0,$$

but for any $x \in \mathbb{R}^n_+$, $G(x) \geq 0$. So for any $\gamma > 0$, the level set $\Lambda_\gamma := \{x : G(x) \leq \gamma\}$ is unbounded, which means the solution set is unbounded if it is not empty. □

From Theorem 3.1, we have the following necessary and sufficient condition for the solution set of ERM$(M(\cdot), q(\cdot))$ to be bounded for any $q(\cdot)$.

THEOREM 3.4. *The solution set of ERM$(M(\cdot), q(\cdot))$ is nonempty and bounded for any $q(\cdot)$ if and only if $M(\cdot)$ is a stochastic R_0 matrix.*

4. Differentiability of G. The objective function G of ERM$(M(\cdot), q(\cdot))$ is, in general, not convex. If G is differentiable at x, then $\min(\nabla G(x), x) = 0$ implies that x is a stationary point of ERM$(M(\cdot), q(\cdot))$. The differentiability of G is studied in [3] for the special case where $M(\omega) \equiv M$, $q(\omega) = \bar{q} + T\omega$ with $M \in \mathbb{R}^{n \times n}$, $\bar{q} \in \mathbb{R}^n$, $T \in \mathbb{R}^{n \times m}$ being constants and T having at least one nonzero element in each row.

In this section, we will give a condition for the function G to be differentiable under a general setting. The continuity of $M(\cdot)$ and $q(\cdot)$ is not assumed.

DEFINITION 4.1. *We say that the strict complementarity condition holds at x with probability one if*

$$P\{\omega : [M(\omega)x]_i + q_i(\omega) = x_i\} = 0, \quad i = 1, \ldots, n.$$

Obviously, this definition is a generalization of the strict complementarity condition for the LCP. The proof for the differentiability of G at x under the strict complementarity condition with probability one is not trivial.

For any fixed ω, if $[M(\omega)x]_i + q_i(\omega) - x_i \neq 0$ for all i, then $\|\Phi(x, \omega)\|^2$ is differentiable at x and

$$\nabla_x \|\Phi(x, \omega)\|^2 = M(\omega)^T (I - D(x, \omega))(M(\omega)x + q(\omega)) + (I + D(x, \omega))x.$$

To simplify the notation, we define

$$f(x, \omega) := M(\omega)^T (I - D(x, \omega))(M(\omega)x + q(\omega)) + (I + D(x, \omega))x. \tag{4.1}$$

THEOREM 4.2. *The function $g(x) := \int_\Omega f(x, \omega) dF(\omega)$ is continuous at x if the strict complementarity condition holds at x with probability one.*

Proof. We will show that $\|g(x+h)-g(x)\| \to 0$ as $h \to 0$. Since

$$\begin{aligned} f(x,\omega)-f(x+h,\omega) &= (M(\omega)^T(I-D(x,\omega))M(\omega)+I+D(x,\omega))h \\ &+ M(\omega)^T(D(x+h,\omega)-D(x,\omega))(M(\omega)(x+h)+q(\omega)) \\ &- (D(x+h,\omega)-D(x,\omega))(x+h), \end{aligned}$$

there exist some constants c_1, $c_2 > 0$ such that

$$\begin{aligned} \|g(x+h)-g(x)\| &= \left\| \int_\Omega [f(x+h,\omega)-f(x,\omega)]dF(\omega) \right\| \\ &\le c_1\|h\| + c_2 \int_\Omega \|D(x+h,\omega)-D(x,\omega)\|dF(\omega). \end{aligned}$$

Then we just need to show that

$$\int_\Omega \|D(x+h,\omega)-D(x,\omega)\|dF(\omega) \to 0 \ \text{ as } \ h \to 0.$$

Note that

$$\{\omega : \|D(x+h,\omega)-D(x,\omega)\| \neq 0\} \subset \cup_{i=1}^n \{A_i \cup B_i\},$$

where

$$A_i := \{\omega : [M(\omega)x]_i + q_i(\omega) - x_i \ge 0,\ [M(\omega)(x+h)]_i + q_i(\omega) - x_i - h_i \le 0\},$$

$$B_i := \{\omega : [M(\omega)x]_i + q_i(\omega) - x_i \le 0,\ [M(\omega)(x+h)]_i + q_i(\omega) - x_i - h_i \ge 0\}.$$

For any $\varepsilon > 0$, since the strict complementarity condition holds at x with probability one, there is a $\delta > 0$ such that

$$P\{\omega : |[M(\omega)x]_i + q_i(\omega) - x_i| < \delta\} < \varepsilon/2. \tag{4.2}$$

Let

$$C_i := \{\omega : [M(\omega)x]_i + q_i(\omega) - x_i \ge \delta,\ [M(\omega)(x+h)]_i + q_i(\omega) - x_i - h_i \le 0\}.$$

Then, we have

$$\begin{aligned} A_i &\subset C_i \cup \{\omega : |[M(\omega)x]_i + q_i(\omega) - x_i| < \delta\}, \\ C_i &\subset \{\omega : [M(\omega)h]_i - h_i \le -\delta\}. \end{aligned}$$

Applying a similar procedure to B_i, we have

$$P\{A_i \cup B_i\} \le P\{\omega : |[M(\omega)h]_i - h_i| \ge \delta\} + P\{\omega : |[M(\omega)x]_i + q_i(\omega) - x_i| < \delta\}.$$

By the Chebyshev inequality, there is an $h_0 > 0$ such that for any h with $\|h\| < h_0$,

$$P\{\omega : |[M(\omega)h]_i - h_i| \ge \delta\} < \varepsilon/2.$$

This together with (4.2) implies that g is continuous at x. ∎

THEOREM 4.3. *If the strict complementarity condition holds at any x in an open set $U \subset \mathbb{R}^n$ with probability one, then G is Fréchet differentiable at $x \in U$ and*

$$\nabla G(x) = \int_\Omega f(x,\omega)dF(\omega). \tag{4.3}$$

Proof. First, we will show that for almost all ω, $\mu\{x \in U : \ [M(\omega)x]_i + q_i(\omega) = x_i\} = 0$ for any i, where μ is Lebesgue measure. If it were not true, then for some i

$$P\{\omega : \mu\{x \in U : \ [M(\omega)x]_i + q_i(\omega) = x_i\} > 0\} > 0,$$

which implies

$$\int_\Omega \int_U 1_{\{[M(\omega)x]_i + q_i(\omega) = x_i\}} dx dF(\omega) > 0. \tag{4.4}$$

But from the assumption and the Fubini theorem [11], we obtain

$$\int_\Omega \int_U 1_{\{[M(\omega)x]_i + q_i(\omega) = x_i\}} dx dF(\omega) = \int_U \int_\Omega 1_{\{[M(\omega)x]_i + q_i(\omega) = x_i\}} dF(\omega) dx = 0.$$

This contradicts (4.4), and hence for almost all ω, $\mu\{x \in U : \ [M(\omega)x]_i + q_i(\omega) = x_i\} = 0$ for any i.

Note that, for any $\omega \in \Omega$, $\|\Phi(x,\omega)\|^2$ is locally Lipschitz and hence absolutely continuous with respect to x. For any (x,ω) such that $[M(\omega)x]_i + q_i(\omega) \neq x_i$, $\|\Phi(x,\omega)\|^2$ is differentiable with respect to x. Therefore, by the Fundamental Theorem of Calculus for Lebesgue Integrals [11], for any x we have

$$\|\Phi(x + h_i e_i, \omega)\|^2 - \|\Phi(x,\omega)\|^2 = \int_0^{h_i} [f(x + s e_i, \omega)]_i ds \tag{4.5}$$

for almost all ω, where $e_i = (0, \ldots, 0, \underset{i}{1}, 0, \ldots, 0)^T$. Thus

$$\begin{aligned} &G(x+h) - G(x) \\ &= \sum_{i=1}^n \int_\Omega \left(\left\| \Phi\left(x + \sum_{k=i}^n h_k e_k, \omega \right) \right\|^2 - \left\| \Phi\left(x + \sum_{k=i+1}^n h_k e_k, \omega \right) \right\|^2 \right) dF(\omega). \end{aligned} \tag{4.6}$$

By (4.5), (4.6), and the Fubini theorem, we deduce that

$$\int_\Omega \int_0^{h_i} [f(y + s e_i, \omega)]_i ds dF(\omega) = \int_0^{h_i} \int_\Omega [f(y + s e_i, \omega)]_i dF(\omega) ds$$

for any i and $y \in B(x, \|h\|) \subset U$, and hence

$$\begin{aligned} &G\,(x+h) - G(x) - h^T \int_\Omega f(x,\omega) dF(\omega) \\ &= \sum_{i=1}^n \int_0^{h_i} \int_\Omega \left[f\left(x + \sum_{k=i+1}^n h_k e_k + s e_i, \omega \right) \right]_i dF(\omega) ds - \sum_{i=1}^n \int_0^{h_i} \int_\Omega [f(x,\omega)]_i dF(\omega) ds \\ &= \sum_{i=1}^n \int_0^{h_i} \int_\Omega \left(\left[f\left(x + \sum_{k=i+1}^n h_k e_k + s e_i, \omega \right) \right]_i - [f(x,\omega)]_i \right) dF(\omega) ds \\ &= \sum_{i=1}^n \int_0^{h_i} \left(g_i\left(x + \sum_{k=i+1}^n h_k e_k + s e_i \right) - g_i(x) \right) ds, \end{aligned}$$

where g is defined in Theorem 4.2. From Theorem 4.2, for any $\varepsilon > 0$, there exists a sufficiently small $h_0 > 0$ such that for any h with $\|h\| < h_0$,

$$\left| G(x+h) - G(x) - h^T \int_\Omega f(x,\omega) dF(\omega) \right| < \varepsilon \|h\|,$$

which implies

$$\frac{|G(x+h) - G(x) - h^T \int_\Omega f(x,\omega) dF(\omega)|}{\|h\|} \to 0 \quad \text{as} \quad \|h\| \to 0.$$

Therefore, G is Fréchet differentiable at x and (4.3) holds. □

Remark. When $M(\omega) \equiv M$ and $q(\omega) = \bar{q} + T\omega$, if $[T\omega]_i$ has no mass at any point for each i, i.e., $P\{\omega : [T\omega]_i = a\} = 0$ for any $a \in \mathbb{R}$, then $P\{\omega : [M(\omega)x]_i + q_i(\omega) = x_i\} = 0$, $i = 1, \ldots, n$. Therefore, if T has at least one nonzero element in each row [3], then for all $x \in \mathbb{R}^n$, the strict complementarity condition holds with probability one and G is differentiable in $\mathbb{R}^n_+$. This indicates that the result shown in Theorem 4.3 contains the results established in [3].

Let $F_{q_i}(s)$ be the distribution function of $q_i(\omega)$, i.e., $F_{q_i}(s) = P\{\omega : q_i(\omega) \le s\}$. Suppose $M(\omega) \equiv M$. Then, we have

$$\begin{aligned} G(x) &= \int_{-\infty}^{+\infty} \sum_{i=1}^n (\min([Mx]_i + s, x_i))^2 dF_{q_i}(s) \\ &= \sum_{i=1}^n \int_{-\infty}^{[(I-M)x]_i} ([Mx]_i + s)^2 dF_{q_i}(s) + \sum_{i=1}^n x_i^2 (1 - F_{q_i}([(I-M)x]_i)). \end{aligned} \tag{4.7}$$

It is shown in [3] that for some special distribution functions, $G(x)$ can be computed without using discrete approximation. The following proposition shows that, under some conditions, we can also compute $\nabla G(x)$ without using discrete approximation.

PROPOSITION 4.4. *If $M(\omega) \equiv M$ and $F_{q_i}(s)$ is a continuous function for all i, then*

$$\nabla G(x) = 2M^T H(x)x + 2(I - H(x))x - 2M^T v(x), \tag{4.8}$$

where

$$H(x) := \operatorname{diag}(F_{q_1}([(I-M)x]_1), \ldots, F_{q_n}([(I-M)x]_n),$$

$$v(x) := \left(\int_{-\infty}^{[(I-M)x]_1} F_{q_1}(s) ds, \ldots, \int_{-\infty}^{[(I-M)x]_n} F_{q_n}(s) ds \right)^T.$$

Proof. If $M(\omega) \equiv M$ and $F_{q_i}(s)$ is continuous for all i, then $P\{\omega : q_i(\omega) = a\} = 0$ for any $a \in \mathbb{R}$, and hence $P\{\omega : [Mx]_i + q_i(\omega) = x_i\} = 0$ for each $x \in \mathbb{R}^n_+$. Then, by Theorem 4.3, $G(x)$ is differentiable at any $x \in \mathbb{R}^n_+$ and

$$\begin{aligned} \nabla G(x) &= \int_\Omega [M^T(I - D(x,\omega))(Mx + q(\omega)) + (I + D(x,\omega))x] dF(\omega) \\ &= M^T \left[\int_\Omega (I - D(x,\omega)) dF(\omega) Mx + \int_\Omega (I - D(x,\omega)) q(\omega) dF(\omega) \right] \\ &\quad + \left(\int_\Omega (I + D(x,\omega)) dF(\omega) \right) x \\ &= 2M^T H(x) Mx + 2M^T R(x) + 2(I - H(x))x, \end{aligned} \tag{4.9}$$

where

$$R(x) := \left(\int_{-\infty}^{[(I-M)x]_1} sdF_{q_1}(s), \ldots, \int_{-\infty}^{[(I-M)x]_n} sdF_{q_n}(s) \right)^T.$$

By integration by parts, we have

$$\int_{-\infty}^{[(I-M)x]_i} sdF_{q_i}(s) = [(I-M)x]_i F_{q_i}([(I-M)x]_i) - \int_{-\infty}^{[(I-M)x]_i} F_{q_i}(s)ds.$$

This implies that

$$R(x) = \left(\int_{-\infty}^{[(I-M)x]_1} sdF_{q_1}(s), \ldots, \int_{-\infty}^{[(I-M)x]_n} sdF_{q_n}(s) \right)^T = H(x)(I-M)x - v(x).$$

Combining this with (4.9), we have the desired formula (4.8). ∎

From (4.8), we see that the smoothness of $G(\cdot)$ depends on the smoothness of $F_{q_i}(\cdot)$, $i = 1, \ldots, n$. If for all i, $F_{q_i}(\cdot)$ is differentiable at $[(I-M)x]_i$ and $\rho_i(\cdot)$, the derivative of $F_{q_i}(\cdot)$ is continuous at $[(I-M)x]_i$, then the Hessian matrix of $G(x)$ can be written as

$$\begin{aligned} \nabla^2 G(x) &= 2M^T H(x)M + 2(M^T S(x) + S(x)M) - 2M^T S(x)M + 2(I - S(x) - H(x)) \\ &= 2M^T H(x)M + 2(I - H(x)) - 2(I-M)^T S(x)(I-M), \end{aligned} \tag{4.10}$$

where

$$S(x) := \operatorname{diag}(x_1\rho_1([(I-M)x]_1), \ldots, x_n\rho_n([(I-M)x]_n)).$$

5. Optimality conditions and error bounds. In numerical algorithms, residual functions play an important role in terminating iterations and verifying accuracy of a computed solution. The following theorem shows the basic properties of the residual function defined by

$$r(x) = \| \min(\nabla G(x), x)\|.$$

THEOREM 5.1. *Suppose that the strict complementarity condition holds at any x in an open set U with probability one. Then the following statements are true.*

(1) *If $\bar{x} \in U$ is a local solution of ERM$(M(\cdot), q(\cdot))$, then $r(\bar{x}) = 0$.*
(2) *If $G(\cdot)$ is twice continuously differentiable at $\bar{x} \in \mathbb{R}^n_+$, where $r(\bar{x}) = 0$ and the Hessian matrix $\nabla^2 G(\bar{x})$ is positive definite, then there are an open set $\bar{U} \subset U$ and a constant $\tau > 0$ such that $\bar{x}$ is a unique local solution of ERM$(M(\cdot), q(\cdot))$ in $\bar{U}$, and for all $x \in \bar{U}$*

$$\|x - \bar{x}\| \le \tau r(x). \tag{5.1}$$

Proof. From Theorem 4.3, we can write the first order optimality condition for the ERM problem (1.3) as $r(x) = 0$. Now we show (5.1). Since $G(\cdot)$ is twice continuously differentiable at $\bar{x}$ and $\nabla^2 G(\bar{x})$ is positive definite, there is an open set $\bar{U} \subset U$ such that $\nabla G(x)$ is a locally Lipschitz continuous and uniform P function in $\bar{U}$. Applying Proposition 6.3.1 in [7] to the nonsmooth equation $\min(x, \nabla G(x)) = 0$, we obtain (5.1). ∎

COROLLARY 5.2. *If $\Omega = \{\omega_1, \ldots, \omega_N\}$ and the strict complementarity condition holds at $x \in \mathbb{R}^n_+$ with probability one, then $r(x) = 0$ implies x is a local solution of ERM$(M(\cdot), q(\cdot))$.*

Proof. Since the strict complementarity condition holds at $x \in \mathbb{R}^n_+$ with probability one, by (3.1), for each i, $\|\Phi(x, \omega_i)\|^2$ is twice continuously differentiable and

$$\nabla^2 G(x) = \sum_{i=1}^{N} [M(\omega_i)^T (I - D(x, \omega_i)) M(\omega_i) + (I + D(x, \omega_i))].$$

Since the Hessian matrix $\nabla^2 G(x)$ is positive semidefinite, and $G(x)$ is a quadratic function in $B(x, \nu)$ for a sufficiently small $\nu > 0$, x is a local solution. □

Now we consider error bounds for the case where G is not necessarily differentiable. Let

$$s(x) = G(x) - \min_{x \in \mathbb{R}^n_+} G(x).$$

When $\Omega = \{\omega_1, \ldots, \omega_N\}$, we can write

$$G(x) = \sum_{j=1}^{N} \sum_{i=1}^{n} |\min([M(\omega_j)x + q(\omega_j)]_i, x_i)|^2 p(\omega_j),$$

where $p(\omega_j)$ is the probability of ω_j. Clearly, there exist finitely many convex polyhedra such that G is a convex quadratic function on each polyhedron, i.e., G is a piecewise convex quadratic function. By Theorem 2.5 in [15], we have the following local error bound result.

PROPOSITION 5.3. *If $\Omega = \{\omega_1, \ldots, \omega_N\}$, then there exist constants $\tau > 0$ and $\varepsilon > 0$ such that for any $x \in \mathbb{R}^n_+$ with $s(x) \le \varepsilon$*

$$\|x - x^*(x)\| \le \tau s(x)^{1/2},$$

where $x^*(x)$ is a global solution of ERM$(M(\cdot), q(\cdot))$ closest to x under the norm $\|\cdot\|$.

Let us denote $s_\gamma(x) = s(x)^\gamma$ for $\gamma > 0$. For a general continuous distribution of ω, G may not be a piecewise convex function. The following example shows that the function s_γ provides a local error bound for the ERM$(M(\cdot), q(\cdot))$ with various values of γ depending on the distribution of ω.

Example 5.1. Consider the SLCP$(M(\omega), q(\omega))$ with

$$M(\omega) \equiv \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & 0 & -1 \end{pmatrix}, \; q(\omega) = \begin{pmatrix} -1 \\ -1 \\ \omega \end{pmatrix},$$

where ω is a random variable with supp$\Omega \subset [-1, 0]$. It is easy to check that $M(\omega)$ is an R_0 matrix. For any ω, the solution set of LCP$(M(\omega), q(\omega))$ is $\{(x_1, 1 - x_1, 0)^T : x_1 \in [-\omega, 1]\} \cup \{(-\frac{1}{2}\omega + \frac{1}{2}, 0, \frac{1}{2}\omega + \frac{1}{2})^T\}$. Let $\rho(\omega)$ be the density function of ω. We consider the following two cases: $\rho(\omega) \equiv 1$ and $\rho(\omega) = 2(\omega + 1)$. Clearly, $x^* = (1, 0, 0)^T$ is the unique global solution of ERM$(M(\cdot), q(\cdot))$ and $r(x^*) = 0$ for these two cases. But for any $x = (x_1, 1 - x_1, 0)^T$ with $x_1 \in [0, 1]$, if $\rho(\omega) \equiv 1$, then $s(x) = (1 - x_1)^{3/2}/\sqrt{3}$, but if $\rho(\omega) = 2(\omega + 1)$, then $s(x) = (1 - x_1)^2/\sqrt{6}$. Noticing that $\|x - x^*\| = \sqrt{2}(1 - x_1)$, we have $\|x - x^*\| \le \tau s_\gamma(x)$, where γ depends on the distribution of ω. So the general

form of local error bound for ERM$(M(\cdot), q(\cdot))$ with continuous random variables is difficult to obtain unless the information on the distribution of ω is known.

THEOREM 5.4. *Let $M(\cdot)$ be a stochastic R_0 matrix. Then for any $\varepsilon > 0$, there exists $\tau > 0$ such that for each $x \in \mathbb{R}^n_+$ with $s(x) > \varepsilon$*

$$\|x - x^*(x)\| \le \tau s(x)^{1/2},$$

where $x^*(x)$ is a global solution of ERM$(M(\cdot), q(\cdot))$ closest to x under the norm $\|\cdot\|$.

Proof. If the assertion were not true, then for any positive integer k, there exists x^k with $s(x^k) > \varepsilon$ such that

$$\|x^k - x^*(x^k)\| > k s(x^k)^{1/2} > k\varepsilon^{1/2}.$$

Since $M(\cdot)$ is a stochastic R_0 matrix, by Theorem 3.4 the global solution set of ERM$(M(\cdot), q(\cdot))$ is nonempty and bounded. Therefore, $\|x^k\| \to \infty$ as $k \to \infty$ and

$$\frac{s(x^k)^{1/2}}{\|x^k\|} \le \frac{\|x^k - x^*(x^k)\|}{k\|x^k\|} \to 0 \text{ as } k \to \infty. \tag{5.2}$$

Let $\{x^{n_k}/\|x^{n_k}\|\}$ be a convergent subsequence of $\{x^k/\|x^k\|\}$. Note that

$$\begin{aligned}\lim_{k\to\infty} \frac{s(x^{n_k})^{1/2}}{\|x^{n_k}\|} &= \lim_{k\to\infty} \frac{(G(x^{n_k}) - \min_{x\in\mathbb{R}^n_+} G(x))^{1/2}}{\|x^{n_k}\|} = \lim_{k\to\infty} \frac{G(x^{n_k})^{1/2}}{\|x^{n_k}\|} \\ &= \lim_{k\to\infty} \left(\int_\Omega \sum_{i=1}^n \left| \min\left(\frac{[M(\omega)x^{n_k}]_i + q_i(\omega)}{\|x^{n_k}\|}, \frac{x_i^{n_k}}{\|x^{n_k}\|} \right) \right|^2 dF(\omega) \right)^{1/2}.\end{aligned}$$

Since for any x with $\|x\| = 1$,

$$\int_\Omega \sum_{i=1}^n |\min([M(\omega)x]_i + q_i(\omega), x_i)|^2 dF(\omega) \le \int_\Omega (\|M(\omega)\|^2 + \|q(\omega)\|^2) dF(\omega) + 1 < \infty,$$

by the dominated convergence theorem, we obtain

$$\lim_{k\to\infty} \frac{s(x^{n_k})^{1/2}}{\|x^{n_k}\|} = \left(\int_\Omega \sum_{i=1}^n |\min([M(\omega)\hat{x}]_i, \hat{x}_i)|^2 dF(\omega) \right)^{1/2} < \infty,$$

where $\hat{x}$ is an accumulation point of $\{x^{n_k}/\|x^{n_k}\|\}$. This, together with (5.2) and $s(x^{n_k})^{1/2}/\|x^{n_k}\| \ge 0$, yields

$$\int_\Omega \sum_{i=1}^n |\min([M(\omega)\hat{x}]_i, \hat{x}_i)|^2 dF(\omega) = 0,$$

which implies that $\hat{x}$ is a solution of the ERM$(M(\cdot), 0)$. Since $\|\hat{x}\| = 1$, this contradicts the assumption that $M(\cdot)$ is a stochastic R_0 matrix from Theorem 2.2 (iii). ∎

Remark. If Ω contains only one element ω and LCP$(M(\omega), q(\omega))$ has a solution, then error bounds in Theorems 5.1 and 5.4 reduce to the local and global error bounds for the R_0 matrix LCP given in [16]. Hence the two theorems are extensions of error bounds for the R_0 matrix LCP given in [16] to the stochastic R_0 matrix LCP in the ERM formulation.

6. Examples and numerical results. In this section, we report numerical results of four examples of the stochastic R_0 matrix LCP in the ERM formulation.

Let the measure of feasibility of $x \in \mathbb{R}^n_+$ with tolerance $\varepsilon \geq 0$ be defined by

$$rel_\varepsilon(x) = P\{\omega : [M(\omega)x]_i + q_i(\omega) \geq -\varepsilon,\ i = 1, \ldots, n\}. \tag{6.1}$$

This measure indicates how much we may expect that x satisfies the constraints $M(\omega)x + q(\omega) \geq 0$ (with some tolerance).

Example 6.1. We use $\tilde{M}$ and $M_0(\omega_0)$ given in Example 2.2, and

$$M_1(\omega') = \begin{pmatrix} 0 & 0 & 0 & -\omega_1 & 0 \\ 0 & 0 & 0 & 0 & -0.4 - 0.4\ln\omega_2 \\ 0 & 0 & 0 & 0 & 0 \\ \omega_1 & 0 & 0 & -2\sqrt{3}\omega_3 & -2\sqrt{3}\omega_3 \\ 0 & 0.4 + 0.4\ln\omega_2 & 0 & -3\omega_4 & 3\omega_4 \end{pmatrix},$$

where $\omega' = (\omega_1, \ldots, \omega_4)^T$ with the distributions of ω_1, ω_2, ω_3, ω_4 being $\mathcal{U}[-0.8, 0.8]$, $\mathcal{U}[0, 1]$, $\mathcal{N}(0, 1)$, and $\mathcal{N}(0, 1)$, respectively. Let $\omega = (\omega_0, \omega_1, \ldots, \omega_4)^T$ and $M(\omega) = \tilde{M} + M_0(\omega_0) + M_1(\omega')$. From Example 2.2, we know that $\tilde{M} + M_0(\omega_0)$ is a stochastic R_0 matrix. It is easy to verify that $E\{M_1(\omega')\} = 0$. Hence by Proposition 2.7, $M(\cdot)$ is a stochastic R_0 matrix.

We set $q(\omega) = \tilde{q} + q_0(\omega)$, where $\tilde{q}$ is a constant vector and $E\{q_0(\omega)\} = 0$. In this example, we choose

$$q_0(\omega) = (0.1\omega_0, 0.1\omega_0, 0, -2\sqrt{3}\omega_3, -3\omega_4)^T$$

with three different cases for $\tilde{q}$,

$$\tilde{q}^1 = (2, 3, 100, -180, -162)^T,\ \tilde{q}^2 = (-5, -5, 0, 10, 10)^T,\ \tilde{q}^3 = (-5, -5, -5, -5, -5)^T.$$

The deterministic LCP$(\tilde{M}, \tilde{q}^i)$, $i = 1, 2, 3$, have a unique solution $(36, 18, 0, 0.25, 0.5)^T$, multiple solutions $(0, 0, \lambda, 0, 0)^T$ with $\lambda \geq 5$, and no feasible solution, respectively.

For all $q_i(\omega)$, we can check that for any $x = (x_1, \ldots, x_5)^T \in \mathbb{R}^5_+$ with $x_i \neq 0, i = 1, 2$, the strict complementarity condition holds at x with probability one and so $\nabla G(x)$ exists at these points. Hence we can use a stochastic approximation algorithm [2, 14, 18] to find a minimizer of $G(x)$ in $\mathbb{R}^n_+$. The iterative formula is given by

$$x^{k+1} = \max(x^k - a_k f(x^k, \omega^k), 0), \tag{6.2}$$

where $f(x, \omega)$ is defined by (4.1), a_k is a stepsize satisfying $\sum_{k=1}^{\infty} a_k = \infty$ and $a_k \to 0$, and ω^k is the kth sample of ω. By the convergence theorems of stochastic approximation algorithms (see [2, Theorem 2.2.1] and [14, Theorem 5.2.1]), the generated sequence $\{x^k\}$ will converge to a connected set S such that every $\bar{x} \in S$ satisfies $\min(g(\bar{x}), \bar{x}) = 0$ with $g(x)$ defined in Theorem 4.2. If $\bar{x}_i \neq 0, i = 1, 2$, then by Theorem 4.3, $\nabla G(\bar{x}) = g(\bar{x})$. In this example, a_k is chosen as

$$a_k = \begin{cases} 0.003, & k \leq 10^4, \\ 0.0025, & 10^4 < k \leq 10^5, \\ 0.002, & 10^5 < k \leq 5 \times 10^5, \\ \frac{1}{k^{0.6}}, & 5 \times 10^5 < k \leq 2 \times 10^6. \end{cases}$$

When $k \geq 5 \times 10^5$, we use the averaging technique proposed by [18] to accelerate the convergence.

TABLE 6.1
Simulation results for Example 6.1 *where* $E\{M(\omega)\}$ *is not an* R_0 *matrix.*

	x_1	x_2	x_3	x_4	x_5	$G(x)$	$r(x)$
$\tilde{q}^1 = (2, 3, 100, -180, -162)^T$							
min	39.5439	23.298	0	0.2079	0.345	7.2405	0.0115
max	40.1396	23.5793	0.0096	0.3804	0.5413	7.5741	0.6486
average	39.7865	23.4563	0.0014	0.2610	0.4635	7.413	0.15826
$\tilde{x}$	36	18	0	0.25	0.5	197.03	12025
$\tilde{q}^2 = (-5, -5, 0, 10, 10)^T$							
min	0.0004	0.0044	11.4092	0	0	1.8518	0
max	0.0030	0.0154	11.6959	0	0	1.9037	0.002
average	0.0008	0.0068	11.5410	0	0	1.8747	4.44×10^{-5}
$\tilde{x}$	0	0	5	0	0	3.1428	0.5718
$\tilde{q}^3 = (-5, -5, -5, -5, -5)^T$							
min	0.004	1.3915	5.7993	0.0005	0.1137	51.652	0.0440
max	0.0377	1.5017	5.896	0.0186	0.2343	51.943	5.2424
average	0.011	1.4347	5.8414	0.003	0.1555	51.734	1.2536

The stochastic approximation algorithm is a local optimization algorithm. To avoid being trapped in a local minimum, for each $\tilde{q}^i$, $i = 1, 2, 3$, we executed 36 times simulation from different initial points $x^0 = (10l, 10l', 0, 0, 0)^T$, $l, l' \in \{0, 1, \ldots, 5\}$. The step size a_k and initial points were chosen based on suggestions for stochastic approximation algorithms in [14].

For each $\tilde{q}^i$, the information on the last iterate x^{kmax}, where $kmax = 2 \times 10^6$, obtained by (6.2) is shown in Table 6.1. The columns labeled as "$G(x)$" and "$r(x)$" show the respective values obtained by the Monte Carlo method with 10^6 samples. The row labeled as "average" shows the average of the values obtained from 36 different initial points. The rows labeled as "min" and "max" indicate the interval of those values, which represents the variability of the values obtained from 36 different initial points.

Recall that the EV method solves the deterministic LCP$(\tilde{M}, \tilde{q})$. Let $\tilde{x}$ be a solution of LCP$(\tilde{M}, \tilde{q})$. For $\tilde{q}^2 = (-5, -5, 0, 10, 10)^T$, since there are multiple solutions $(0, 0, \lambda, 0, 0)^T$ with $\lambda \geq 5$, $G(x)$, and $r(x)$ are evaluated at $\tilde{x} = (0, 0, 5, 0, 0)^T$. There is no feasible solution of LCP$(\tilde{M}, \tilde{q})$ with $\tilde{q}^3 = (-5, -5, -5, -5, -5)^T$.

By using the Monte Carlo method with 10^6 samples, we evaluated the measure of feasibility rel_ε defined by (6.1) for the case of $\tilde{q}^1 = (2, 3, 100, -180, -162)^T$, at $\tilde{x} = (36, 18, 0, 0.25, 0.5)^T$ and at the last iterates obtained by the iterative formula (6.2) from 36 different initial points. The results are presented in Table 6.2. The row labeled as "ave.rel_ϵ" shows the average values of $rel_\varepsilon(x^{kmax})$ obtained at the 36 last iterates x^{kmax}. For each $\tilde{q}^i$ and initial points, the computational time for obtaining the values in Table 6.1 is about 185 seconds by MATLAB 7.0 with a computer with a Pentium 4 3.06 GHz CPU.

Example 6.2. In this example, we consider the case where $M(\omega) \equiv \tilde{M}$ is a P matrix and $q(\omega)$ has continuous distribution. In this case, the EV formulation LCP$(\tilde{M}, \tilde{q})$ has a unique solution $\tilde{x}$. The objective function G of the ERM formulation is twice continuously differentiable and the values of $G(x)$, $\nabla G(x)$, and $\nabla^2 G(x)$ can be computed by (4.7), (4.8), (4.10), respectively, without resorting to stochastic approximation.

Let $q(\omega) = \tilde{q} + q_0(\omega)$, where $\tilde{q} = E\{q(\omega)\}$, $q_0(\omega) = B\omega$, $\omega = (\omega_1, \omega_2, \omega_3)^T \in$

TABLE 6.2
$rel_\varepsilon(\tilde{x})$ and average of $rel_\varepsilon(x^{kmax})$ for Example 6.1 with $\tilde{q}^1$ in Table 6.1.

ε	0.0	0.1	0.2	0.5	1
$rel_\epsilon(\tilde{x})$	0.0018	0.0581	0.2971	0.3236	0.3417
ave.$rel_\epsilon(x^{kmax})$	0.3084	0.7190	0.9007	0.9488	0.9518

TABLE 6.3
Simulation results for Example 6.2 where $E\{M(\omega)\}$ is a P matrix.

	$n = 20$		$n = 50$		$n = 100$	
	$r(x)$	$G(x)$	$r(x)$	$G(x)$	$r(x)$	$G(x)$
$\tilde{x}$	97.39	180.77	168.04	427.58	307.42	823.8
$\bar{x}$	6.17×10^{-8}	75.78	5.14×10^{-8}	167.07	1.34×10^{-7}	293.72
$\|\tilde{x} - \bar{x}\|$	7.51		21.03		38.18	

TABLE 6.4
$rel_\varepsilon(\tilde{x})$ and $rel_\varepsilon(\bar{x})$ for Example 6.2.

	$n = 20$			$n = 50$			$n = 100$		
ε	0	1	5	0	1	5	0	1	5
$rel_\varepsilon(\tilde{x})$	0.2507	0.3387	0.6615	0.2072	0.2930	0.6152	0.1856	0.2709	0.5934
$rel_\varepsilon(\bar{x})$	0.3863	0.4955	0.8049	0.2712	0.3713	0.7225	0.2208	0.3193	0.6723

$\mathcal{N}(0, I)$, $B \in \mathbb{R}^{n\times 3}$ is 100% dense, and the elements of B are randomly generated with the uniform distribution $\mathcal{U}(0, 5)$. We use Example 4.4 of [4] to generate $\tilde{M}$ and $\tilde{q}$. First, we randomly generate 100% dense $A \in \mathbb{R}^{n\times n}$ and $\bar{q} \in \mathbb{R}^n$ whose elements are uniformly distributed in $(-5, 5)$. Then we use the QR decomposition of A to get an upper triangular matrix N, and obtain a triangular matrix $\tilde{M}$ by replacing the diagonal elements of N by their absolute values.

We first use Lemke's method [8] to find a solution $\tilde{x}$ of LCP$(\tilde{M}, \tilde{q})$, and then take $\tilde{x}$ as an initial point to find a local solution $\bar{x}$ of the ERM formulation by applying the semismooth Newton method [7] to the equation $\min(\nabla G(x), x) = 0$.

The numerical experiments were carried out for $n = 20$, 50, and 100. For each n, we generated 100 problems and solved them by the above-mentioned procedure. The figures presented in Table 6.3 are the average of the results obtained in this manner.

The measures rel_ε of feasibility at $\tilde{x}$ and $\bar{x}$ obtained by the EV method and the ERM method, respectively, are presented in Table 6.4.

Example 6.3. To illustrate the application of stochastic R_0 matrix linear complementarity problems, we use a simple transportation network shown in Figure 6.1, which is based on an example of the deterministic traffic equilibrium network model in [6].

In the network, two cities West and East are connected by two two-way roads and one one-way road. More specifically, the network consists of five links, L1, L2, L3, R1, R2, where L1, L2, L3 are directed from West to East, and R1 and R2 are the returns of L1 and L2, respectively. L1-R1 is a mountain road, and L2-R2 and L3 are seaside roads. We are interested in the traffic flow between the two cities. The Wardrop equilibrium principle states that each driver will choose the minimum cost route between the origin-destination pair, and through this process the routes that are used will have equal cost; routes with costs higher than the minimum will have no flow. In a deterministic model, the parameters in the demand and cost function are fixed, and the problem can be formulated as a (deterministic) LCP based on the

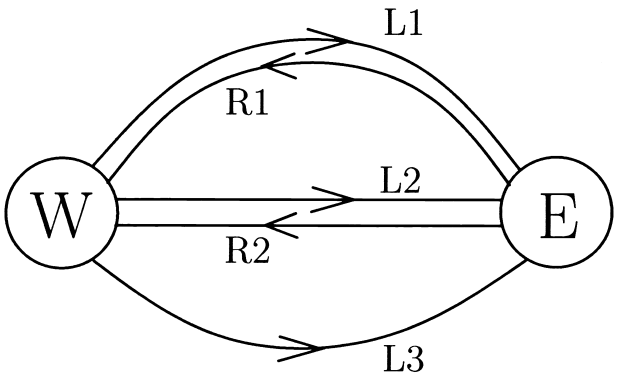


FIG. 6.1. *Road network.*

Wardrop equilibrium principle.

In practice, however, the traffic condition will be significantly affected by some uncertain factors such as weather. So we want to estimate the traffic flow and the travel time that are most likely to occur, before we know such uncertain factors.[1]

We suppose that there are three possible uncertain weather conditions; sunny, windy, and rainy. On a sunny day, the network is free from traffic congestion and the travel times of all roads are constant, which are given as $(c_1, c_2, c_3, c_4, c_5)^T = (1000, 950, 3000, 1000, 1300)^T$, where c_1, c_2, c_3, c_4, c_5 denote the travel times of roads L1, L2, L3, R1, R2, respectively. On a windy day, the seaside roads suffer from traffic jams due to congestion, and the travel times of the roads in the whole network are given by

$$\begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 60 & 0 & 0 & 20 \\ 0 & 0 & 80 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 & 100 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \end{pmatrix} + \begin{pmatrix} 1000 \\ 950 \\ 3000 \\ 1000 \\ 1300 \end{pmatrix},$$

where v_1, v_2, v_3, v_4, v_5 denote the traffic volumes of roads L1, L2, L3, R1, R2, respectively. On the other hand, on a rainy day the mountain roads suffer from traffic jams, and the travel times of the roads in the whole network are given by

$$\begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \end{pmatrix} = \begin{pmatrix} 40 & 0 & 0 & 20 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 8 & 0 & 0 & 80 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \end{pmatrix} + \begin{pmatrix} 1000 \\ 950 \\ 3000 \\ 1000 \\ 1300 \end{pmatrix}.$$

Moreover, trip demands between the two cities are higher on a sunny day than on a windy or rainy day. Specifically, $(d_1, d_2)^T = (260, 170)^T$ on a sunny day and $(d_1, d_2)^T = (160, 70)^T$ on a windy or rainy day, where d_1 and d_2 are trip demands from West to East and from East to West, respectively.

It is convenient to represent the travel cost functions and trip demands in a unified

[1]It should be noted that we do not intend to construct a traffic equilibrium model in which the drivers choose their routes under uncertainty.

manner as follows:

$$c(v,\omega) = H(\omega)v + h,$$

where

$$c(v,\omega) = (c_1(v,\omega),\ldots,c_5(v,\omega))^T,$$

$$H(\omega) = \begin{pmatrix} 40\alpha(\omega) & 0 & 0 & 20\alpha(\omega) & 0 \\ 0 & 60\beta(\omega) & 0 & 0 & 20\beta(\omega) \\ 0 & 0 & 80\beta(\omega) & 0 & 0 \\ 8\alpha(\omega) & 0 & 0 & 80\alpha(\omega) & 0 \\ 0 & 4\beta(\omega) & 0 & 0 & 100\beta(\omega) \end{pmatrix},$$

$$\alpha(\omega) = \frac{1}{2}\omega(\omega-1), \quad \beta(\omega) = \omega(2-\omega), \quad h = (1000, 950, 3000, 1000, 1300)^T.$$

Here $\Omega = \{\omega^1, \omega^2, \omega^3\}$ with $\omega^1 = 0$, $\omega^2 = 1$, $\omega^3 = 2$ represents the set of uncertain events of the weather, {sunny, windy, rainy}, with probabilities $p_1 = \frac{1}{2}$, $p_2 = \frac{1}{4}$, $p_3 = \frac{1}{4}$, respectively. Also, the traffic flow $v = (v_1, v_2, v_3, v_4, v_5)^T$ should satisfy[2]

$$v \geq 0, \quad Bv \geq d(\omega),$$

where

$$B = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}, \quad d(\omega) = \begin{pmatrix} 260 - 100(\alpha(\omega) + \beta(\omega)) \\ 170 - 100(\alpha(\omega) + \beta(\omega)) \end{pmatrix}.$$

By Wardrop's principle, for each event $\omega \in \Omega$, the traffic equilibrium problem can be formulated as LCP$(M(\omega), q(\omega))$ with

$$M(\omega) = \begin{pmatrix} H(\omega) & -B^T \\ B & 0 \end{pmatrix}, \; q(\omega) = \begin{pmatrix} h \\ -d(\omega) \end{pmatrix}.$$

The solutions $x(\omega^i)$ of LCP$(M(\omega^i), q(\omega^i))$, $i = 1, 2, 3$ express the equilibrium traffic flow on each link as well as the minimum travel time between each origin-destination pair, on a sunny day, a windy day, and a rainy day, respectively. The average traffic flow is given by $(E\{x_1(\omega)\},\ldots,E\{x_5(\omega)\})$, and the average travel time on each direction is given by $(E\{x_6(\omega)\}, E\{x_7(\omega)\})$.

On the other hand, the average travel costs and demands are given by

$$E\{c(v,\omega)\} = \begin{pmatrix} 10 & 0 & 0 & 5 & 0 \\ 0 & 15 & 0 & 0 & 5 \\ 0 & 0 & 20 & 0 & 0 \\ 2 & 0 & 0 & 20 & 0 \\ 0 & 1 & 0 & 0 & 25 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \end{pmatrix} + \begin{pmatrix} 1000 \\ 950 \\ 3000 \\ 1000 \\ 1300 \end{pmatrix},$$

$$E\{d(\omega)\} = \begin{pmatrix} 210 \\ 120 \end{pmatrix},$$

[2]For the purpose of our presentation, we may replace the equality constraint $Bv = d(\omega)$ by the inequality constraint. In practice, this change will not affect the solution of the problem.

TABLE 6.5
Traffic flow and travel time for Example 6.3.

EV solution $\tilde{x}$	(120, 90, 0, 70, 50, 2550, 2640)
ERM solution $\bar{x}$	(84, 84, 21, 80, 20, 975, 1000)
$x(\omega^1)$	(0, 260, 0, 170, 0, 950, 1000)
$x(\omega^2)$	(955/6, 5/6, 0, 70, 0, 1000, 1000)
$x(\omega^3)$	(0, 160, 0, 3.75, 66.75, 950, 1300)
$E\{x(\omega)\}$	(39.8, 170.2, 0, 103.4, 16.6, 962.5, 1075)
$E\{\|\|x(\omega)-\tilde{x}\|\|\}$, $\|E\{x(\omega)\}-\tilde{x}\|\|$	2239.66, 2232.60
$E\{\|\|x(\omega)-\bar{x}\|\|\}$, $\|E\{x(\omega)\}-\bar{x}\|\|$	222.42, 127.16

which are exactly the same as those of the five-link example in [6].

Below we compare the estimates of the traffic flows and travel time obtained by the EV formulation and the ERM formulation.

The solution of the EV formulation, LCP$(E\{M(\omega)\}, E\{q(\omega)\})$, is denoted by $\tilde{x}$. The ERM formulation for this example is the problem of minimizing the function

$$G(x) = \sum_{i=1}^{3} p_i || \min(x, M(\omega^i)x + q)||^2.$$

We denote the solution by $\bar{x}$. In Table 6.5, we report numerical results.

It is observed from $\tilde{x}_3 = x(\omega^i) = 0, i = 1, 2, 3$ in Table 6.5 that the user-optimal load pattern estimated from the EV formulation has no flow on L3, which is the same as the user-optimal traffic pattern estimated from the LCPs for a sunny day, a windy day, and a rainy day, respectively. However, the estimated total travel time $\tilde{x}_6 + \tilde{x}_7 = 5190$ from the EV formulation is larger than the total travel time obtained from the LCP for any day. On the other hand, the user-optimal traffic pattern estimated from the ERM formulation has light flow on L3 and the total travel time $\bar{x}_6 + \bar{x}_7 = 1975$ is close to $x_6(\omega^i) + x_7(\omega^i), i = 1, 2, 3$.

The two formulations yield different estimates of the user-optimal traffic pattern and travel time, and both solutions, $\bar{x}$ and $\tilde{x}$, try to explain the phenomenon in the real world. The EV formulation uses the average of data to estimate the user-optimal traffic pattern. The ERM formulation uses the least square method to find a traffic pattern which has minimum total error to each user-optimal traffic pattern for each day. It is worth mentioning that, as far as this example is concerned, $\bar{x}$ may be considered closer to the realized traffic patterns than $\tilde{x}$ because $E\{\|x(\omega) - \tilde{x}\|\} > E\{\|x(\omega) - \bar{x}\|\}$ and $\|E\{x(\omega)\} - \tilde{x}\| > \|E\{x(\omega)\} - \bar{x}\|$.

Now, we use this example to show that the theoretical results given in this paper substantially extend the results in [3]. It is easy to verify that the matrix $E\{M(\omega)\}$ is an R_0 matrix. By Proposition 2.5, $M(\cdot)$ is a stochastic R_0 matrix. Hence by Theorem 3.1 the solution set of ERM$(M(\cdot), q(\cdot))$ is nonempty and bounded. However, for each ω^i, $M(\omega^i)$ is not an R_0 matrix. Hence the statement on the solution set cannot be obtained by using the results in [3].

Example 6.4. The last example is a simplified control problem: Let $\hat{\omega} \in \mathbb{R}^n$ be the system parameter. Based on prior experience, we assume that $\hat{\omega}$ is generated from $\mathcal{N}(a, B)$. At each time t, we have the following observer:

$$y_{t+1} = X_t\hat{\omega} + F_t v_t, \tag{6.3}$$

where $X_t \in \mathbb{R}^{m\times n}$ is a known input, $F_t \in \mathbb{R}^{m\times r}$ is a known matrix, and $v_t \in \mathbb{R}^r$ is an unknown noise which is independent identically and normally distributed with $E\{v_t\} = 0$, $E\{v_t v_t^T\} = I$.

Suppose $B \succ 0$ and $F_t F_t^T \succeq 0$. By the Kalman filter theory [1], we have the following recursive estimation for the parameter $\hat{\omega}$:

$$\begin{aligned} \omega_{t+1} &= \omega_t + K_{t+1}(y_{t+1} - X_t\omega_t), \\ K_{t+1} &= B_t X_t^T (X_t B_t X_t^T + F_t F_t^T)^+, \\ B_{t+1} &= B_t - B_t X_t^T K_{t+1}^T, \\ \omega_0 &= a, \ B_0 = B, \end{aligned} \tag{6.4}$$

where A^+ denotes the pseudoinverse of matrix A. Then the posterior distribution of $\hat{\omega}$ is given by $N(\omega_t, B_t)$. The control law u_t is obtained as a solution of the following convex quadratic program:

$$\begin{aligned} \min \ & c_t(\hat{\omega})^T u + \frac{1}{2} u^T Q_t(\hat{\omega}) u \\ \text{s.t.} \quad & A_t(\hat{\omega}) u \le b_t(\hat{\omega}) \\ & u \ge 0, \end{aligned} \tag{6.5}$$

where $Q_t(\hat{\omega})$, $A_t(\hat{\omega})$ are matrices and $c_t(\hat{\omega})$, $b_t(\hat{\omega})$ are vectors. The first order optimality condition of (6.5) is equivalent to the LCP$(M_t(\hat{\omega}), q_t(\hat{\omega}))$ with

$$M_t(\hat{\omega}) = \begin{pmatrix} Q_t(\hat{\omega}) & A_t(\hat{\omega})^T \\ -A_t(\hat{\omega}) & 0 \end{pmatrix}, \ q_t(\hat{\omega}) = \begin{pmatrix} c_t(\hat{\omega}) \\ b_t(\hat{\omega}) \end{pmatrix}.$$

In traditional adaptive control, we replace the unknown parameter $\hat{\omega}$ by its estimate ω_t in the quadratic program (6.5) to obtain an approximation $\check{u}_t$ of the control law u_t for each t, that is, $\check{u}_t$ is the vector whose elements are the first n components of the solution of the LCP$(M_t(\omega_t), q_t(\omega_t))$.

If ω_t is far away from the parameter $\hat{\omega}$, the error of $\check{u}_t$ is big and will cause trouble in some situations. Hence we take the variance of the estimate into account by using the solution $\bar{u}_t$ of the ERM formulation for SLCP$(M_t(\omega), q_t(\omega))$ with $\omega \sim \mathcal{N}(\omega_t, B_t)$. Here we report numerical results for a tracking problem with the ARX model $y_{t+1} = \hat{\omega}^{(1)} y_t + \hat{\omega}^{(2)} u_t + v_t$. The controller u_t would be designed so that y_{t+1} can track a given trajectory $\exp(0.5t)$. Let the performance function be $p(u_t, \omega) := (\omega^{(1)} y_t + \omega^{(2)} u_t - \exp(0.5t))^2$. Then, from

$$p(u_t, \omega) = (\omega^{(2)})^2 u_t^2 - 2(\exp(0.5t) - \omega^{(1)} y_t)\omega^{(2)} u_t + (\exp(0.5t) - \omega^{(1)} y_t)^2,$$

we have $c_t(\omega) = -2(\exp(0.5t) - \omega^{(1)} y_t)\omega^{(2)}$, $Q_t(\omega) = 2(\omega^{(2)})^2$. We set $X_t = (y_t, u_t)$ and choose $a = (0, 1)^T$, $B = \left(\begin{smallmatrix} 0.25 & 0 \\ 0 & 4 \end{smallmatrix}\right)$, $F_t = 1$, $A_t(\omega) \equiv 1$, $b_t(\omega) = 4 + 2(\omega^{(2)})^2$.

For $k \ge 1$, we generate a true parameter $\hat{\omega}^k$ from $\mathcal{N}(1, 1)$ and noise $\{v_t\}$ from $\mathcal{N}(0, 1)$. We solve the ERM formulation for SLCP$(M_t(\omega), q_t(\omega))$ with $\omega \sim \mathcal{N}(\omega_t, B_t)$ to obtain $\bar{u}_t^k$. We then set $X_t = (y_t^k, \bar{u}_t^k)$ and use (6.3) and (6.4) to obtain y_{t+1}^k, ω_{t+1}, and B_{t+1}. We also solve LCP$(M_t(\omega_t), q_t(\omega_t))$ and the EV formulation of SLCP$(M_t(\omega), q_t(\omega))$ with $\omega \sim \mathcal{N}(\omega_t, B_t)$ to get $\check{u}_t^k$ and $\tilde{u}_t^k$, respectively.

TABLE 6.6
Average performance for Example 6.4.

t	1	2	3	4	5
$\bar{\sigma}_t$	1.0103	1.9764	2.286	1.6755	1.8693
$\check{\sigma}_t$	1.7053	3.0257	2.3345	1.7385	1.8918
$\tilde{\sigma}_t$	1.2425	2.5505	2.3852	1.8015	2.0903

For the purpose of comparison, we define the average performance (for $k = 1, 2, \ldots, 100$) of these formulations by

$$\bar{\sigma}_t := \frac{1}{100}\sum_{k=1}^{100}(\bar{u}_t^k - u_t^*(\hat{\omega}^k))^2, \quad t = 1, 2, 3, 4, 5,$$

$$\check{\sigma}_t := \frac{1}{100}\sum_{k=1}^{100}(\check{u}_t^k - u_t^*(\hat{\omega}^k))^2, \quad t = 1, 2, 3, 4, 5,$$

$$\tilde{\sigma}_t := \frac{1}{100}\sum_{k=1}^{100}(\tilde{u}_t^k - u_t^*(\hat{\omega}^k))^2, \quad t = 1, 2, 3, 4, 5,$$

where $u_t^*(\hat{\omega}^k)$ is obtained by solving $\mathrm{LCP}(M_t(\hat{\omega}^k), q_t(\hat{\omega}^k))$ with true parameter $\hat{\omega}^k$.

From the results shown in Table 6.6, we find that the ERM formulation has better performance than $\mathrm{LCP}(M_t(\omega_t), q_t(\omega_t))$ and the EV formulation in the sense that $\bar{\sigma}_t < \check{\sigma}_t$ and $\bar{\sigma}_t < \tilde{\sigma}_t$ hold for all $t = 1, 2, 3, 4, 5$. This suggests that $\bar{u}_t^k$ is a better control law for $\hat{\omega}^k$ than $\tilde{u}_t^k$ and $\bar{u}_t^k$ in all cases.

7. Final remark. This paper proves that a necessary and sufficient condition for the $\mathrm{ERM}(M(\cdot), q(\cdot))$ having a nonempty and bounded solution set is that $M(\cdot)$ is a stochastic R_0 matrix. Proposition 2.5 shows that if the matrix $E\{M(\omega)\}$ is an R_0 matrix, then $M(\cdot)$ is a stochastic R_0 matrix. Moreover, Example 6.1 shows that there are many cases where $M(\cdot)$ is a stochastic R_0 matrix, but $E\{M(\omega)\}$ is not an R_0 matrix, and the EV formulation $\mathrm{LCP}(E\{M(\omega)\}, E\{q(\omega)\})$ either has no solution or has an unbounded solution set. Therefore, the condition for the $\mathrm{ERM}(M(\cdot), q(\cdot))$ having a nonempty and bounded solution set is weaker than the condition for the EV formulation having a nonempty and bounded solution set. Furthermore, when $\mathrm{ERM}(M(\cdot), q(\cdot))$ has a solution $\bar{x}$ and $\mathrm{LCP}(E\{M(\omega)\}, E\{q(\omega)\})$ has a solution $\tilde{x}$, the residuals always satisfy

$$G(\bar{x}) = E\{\|\min(M(\omega)\bar{x} + q(\omega), \bar{x})\|^2\} \le E\{\|\min(M(\omega)\tilde{x} + q(\omega), \tilde{x})\|^2\} = G(\tilde{x}).$$

Example 6.1 shows that $G(\bar{x})$ can be much smaller than $G(\tilde{x})$. Moreover, the values of rel_ε shown in Table 6.2 reveal that, for each tolerance level $\varepsilon \ge 0$, the number of ω^i at which $M(\omega^i)\bar{x} + q(\omega^i) < -\varepsilon$ holds is much less than the number of ω^i at which $M(\omega^i)\tilde{x} + q(\omega^i) < -\varepsilon$ holds. Example 6.2 shows that a local solution $\bar{x}$ of $\mathrm{ERM}(M(\cdot), q(\cdot))$ may conveniently be obtained from a solution $\tilde{x}$ of the EV formulation. Examples 6.3 and 6.4 show that the EV formulation and ERM formulation express different concerns in our real life. Solving the EV formulation is usually less expensive computationally than solving the ERM formulation. Nevertheless, since $\bar{x}$ is generally expected to have better reliability than $\tilde{x}$, we may recommend the ERM method to those decision makers who do not want to take the high risk of violating the conditions $M(\omega)x + q(\omega) \ge 0$.

Acknowledgments. We are very grateful to Professor Jong-Shi Pang and the referees of this paper for their helpful comments.

REFERENCES

[1] P. E. Caines, *Linear Stochastic Systems*, John Wiley & Sons, New York, 1988.
[2] H.-F. Chen, *Stochastic Approximation and Its Applications*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2002.
[3] X. Chen and M. Fukushima, *Expected residual minimization method for stochastic linear complementarity problems*, Math. Oper. Res., 30 (2005), pp. 1022–1038.
[4] X. Chen and Y. Ye, *On homotopy-smoothing methods for box-constrained variational inequalities*, SIAM J. Control Optim., 37 (1999), pp. 589–616.
[5] R. W. Cottle, J. S. Pang, and R. E. Stone, *The Linear Complementarity Problem*, Academic Press, Inc., Boston, 1992.
[6] S. Dafermos, *Traffic equilibrium and variational inequalities*, Transportation Sci., 14 (1980), pp. 42–54.
[7] F. Facchinei and J. S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Vols. I *and* II, Springer-Verlag, New York, 2003.
[8] M. C. Ferris, ftp://ftp.cs.wisc.edu/math-prog/matlab/lemke.m., Department of Computer Sciences, University of Wisconsin, Madison, WI, 1998.
[9] M. C. Ferris and J. S. Pang, *Engineering and economic applications of complementarity problems*, SIAM Rev., 39 (1997), pp. 669–713.
[10] A. Fischer, *A special Newton-type optimization method*, Optimization, 24 (1992), pp. 269–284.
[11] G. B. Folland, *Real Analysis: Modern Techniques and Their Applications*, John Wiley & Sons, New York, 1984.
[12] G. Gürkan, A. Y. Özge, and S. M. Robinson, *Sample-path solution of stochastic variational inequalities*, Math. Program., 84 (1999), pp. 313–333.
[13] P. Kall and S. W. Wallace, *Stochastic Programming*, John Wiley & Sons, Chichester, UK, 1994.
[14] H. J. Kushner and G. Yin, *Stochastic Approximation Algorithms and Applications*, Springer-Verlag, New York, 1997.
[15] W. Li, *Error bounds for piecewise convex quadratic programs and applications*, SIAM J. Control Optim., 33 (1995), pp. 1510–1529.
[16] O. L. Mangasarian and J. Ren, *New improved error bounds for the linear complementarity problem*, Math. Program., 66 (1994), pp. 241–255.
[17] J. S. Pang, *Error bounds in mathematical programming*, Math. Program., 79 (1997), pp. 299–332.
[18] B. T. Polyak and A. B. Juditsky, *Acceleration of stochastic approximation by averaging*, SIAM J. Control Optim., 30 (1992), pp. 838–855.

SIAM J. OPTIM.
Vol. 18, No. 2, pp. 507–536

PEELING OFF A NONCONVEX COVER OF AN ACTUAL CONVEX PROBLEM: HIDDEN CONVEXITY*

Z. Y. WU†, D. LI‡, L. S. ZHANG§, AND X. M. YANG¶

Abstract. Convexity is, without a doubt, one of the most desirable features in optimization. Many optimization problems that are nonconvex in their original settings may become convex after performing certain equivalent transformations. This paper studies the conditions for such hidden convexity. More specifically, some transformation-independent sufficient conditions have been derived for identifying hidden convexity. The derived sufficient conditions are readily verifiable for quadratic optimization problems. The global minimizer of a hidden convex programming problem can be identified using a local search algorithm.

Key words. convexity, hidden convexity, hidden-convex function, hidden-convex programming problem, global optimization

AMS subject classification. 90C30

DOI. 10.1137/050648584

1. Introduction.

We consider the following optimization problem:

$$(P) \qquad \begin{aligned} \min \ & g_0(x) \\ \text{s.t.} \ & g_k(x) \le b_k, k = 1, \ldots, m, \\ & x \in X, \end{aligned}$$

where $g_k : R^n \to R,\ k = 0, 1, \ldots, m,$ are second-order continuously differentiable functions and

$$X = \{x \in R^n \mid l_i \le x_i \le u_i,\ i = 1, \ldots, n\}. \tag{1.1}$$

Without loss of generality, we assume in this paper that $0 < l_i < u_i$ for all $i = 1, \ldots, n$.

When all of the functions g_k, $k = 0, 1, \ldots, m$, are convex, problem (P) is a convex programming problem that can be solved efficiently with many existing algorithms.

Convexity plays a central role in optimization theory. In addition to the many desirable properties that are enjoyed by convexity, convexity guarantees a local minimum to be a global solution at the same time. Observations and experiences of optimization, however, often reveal that convexity in many situations is a property that is associated with a given representation space. More specifically, an equivalent transformation may convert a nonconvex problem in its original setting to a convex problem in a transformed space. In this sense, convexity could be hidden, if the

*Received by the editors December 28, 2005; accepted for publication (in revised form) December 18, 2006; published electronically May 29, 2007. The work was partially supported by the Research Grants Council of Hong Kong under grant CUHK 4180/03E and grant 2050291.

http://www.siam.org/journals/siopt/18-2/64858.html

†Department of Mathematics, Chongqing Normal University, Chongqing 400047, People's Republic of China (zhiyouwu@263.net); current address: School of Information Technology and Mathematical Sciences, University of Ballarat, Ballarat 3353, Victoria, Australia (z.wu@ballarat.edu.au).

‡Corresponding author. Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong (dli@se.cuhk.edu.hk).

§Department of Mathematics, Shanghai University, Shanghai 200436, People's Republic of China (lszhang@staff.shu.edu.cn).

¶Department of Mathematics, Chongqing Normal University, Chongqing 400047, People's Republic of China (xmyang@cqnu.edu.cn).

representation space is "inappropriate." The purpose of this paper is to investigate the conditions for identifying such seemingly nonconvex problems that are actually hidden convex.

It is worth mentioning that convexity has been extended to various forms of generalized convexity in the literature [2], [3], [5], [10], [14], [15]. Examples of generalized convexities include pseudoconvexity and quasiconvexity. For many years, researchers have been exploring the situations in which the convexity condition can be relaxed to a certain degree, while, at the same time, some desirable properties that are similar to those enjoyed by convex functions are preserved.

Horst [8] discussed the concept of range and domain transformations that can convexify some nonconvex functions. Li et al. [12] derived specific sufficient conditions under which some nonconvex functions can be convexified by a domain transformation; especially, the relationship between monotonicity and domain transformation has been recently discussed in [11], [12], [16]. This paper considers specific sufficient conditions under which certain nonconvex functions can be convexified by using a range transformation or using both domain and range transformations. The results are then applied to the study of hidden-convex programming problems.

While the hidden convexity that can be identified by a domain transformation is confined to monotone functions [12], the hidden convexity that can be identified by a range transformation is limited to a subset of pseudoconvex functions. The most interesting finding of this paper is that a combined domain and range transformation is capable of identifying a general class of hidden-convex functions that goes beyond the class of second order differentiable quasiconvex functions. It is also worth pointing out that there is a significant difference between the methodologies involved in proving a hidden convexity only by a domain transformation [12] and those used in revealing a hidden convexity by a combined range and domain transformation. The identification of a general class of hidden-convex functions that is reported in this paper is accomplished by introducing a new concept of the constrained minimum eigenvalue of the Hessian of a transformed function, which requires a more complicated analysis than does the traditional approach of checking the minimum eigenvalue of the Hessian of a transformed function, as was used in [12] for a domain transformation.

This paper is organized as follows. In section 2, we derive some results on constrained minimum eigenvalues that are essential in later derivation for hidden convexity. Hidden convexity is introduced in section 3, and sufficient conditions are developed to recognize hidden-convex functions. The relationships among hidden convexity, pseudoconvexity, quasiconvexity, and monotonicity are discussed. Section 4 extends the earlier investigation to sufficient conditions for hidden-convex programming problems. The most promising results of this paper appear in section 5, in which implementable sufficient conditions are obtained for the identification of hidden-convex quadratic optimization problems. The paper finally concludes in section 6 with some discussion of future research topics.

2. Preliminaries.

2.1. Minimum eigenvalue and constrained minimum eigenvalue. Let $X \subset R^n$ be a compact set and S^n be the unit sphere in R^n, $\{d \in R^n \mid \|d\| = 1\}$. Let $A(x) = (a_{i,j}(x))_{n\times n}$ be a symmetric matrix defined on X with each $a_{i,j}(x)$ being assumed to be continuous and

$$b(x) = (b_1(x), \ldots, b_n(x))^T \tag{2.1}$$

be an n-dimensional vector function defined on X with each $b_i(x)$ being assumed to be continuous. We define the following for given matrix $A(x)$, vector $b(x)$, and a positive number q:

$$\lambda = \min_{x\in X, d\in S^n} d^T A(x) d, \tag{2.2}$$

$$\gamma = \min_{x\in X, d\in S^n, b^T(x)d=0} d^T A(x) d, \tag{2.3}$$

$$B_q(x) = qb(x)b^T(x) + A(x),$$

$$\mu_q = \min_{x\in X, d\in S^n} d^T B_q(x) d.$$

LEMMA 2.1. *For any $q > 0$,*

$$\lambda \le \mu_q \le \gamma$$

and

$$\lim_{q\to+\infty} \mu_q = \gamma.$$

Proof. The first part of the lemma is obvious. Therefore,

$$\limsup_{q\to+\infty} \mu_q \le \gamma. \tag{2.4}$$

For the second part of the lemma, we thus only need to prove that

$$\liminf_{q\to+\infty} \mu_q \ge \gamma. \tag{2.5}$$

Suppose that for any k, there exist $q_k > k$, $d_k \in S^n$, and $x_k \in X$, such that

$$d_k^T \Big(q_k b(x_k) b^T(x_k) + A(x_k) \Big) d_k \le \gamma - \alpha_0, \tag{2.6}$$

where α_0 is a positive number.

Because S^n and X are all compact sets, there must exist a subsequence $\{k_j\}$ of $\{k\}$, such that

$$\lim_{j\to\infty} d_{k_j} = d_0, \quad \lim_{j\to\infty} x_{k_j} = x_0,$$

where $d_0 \in S^n$ and $x_0 \in X$. We thus have the following from (2.6):

$$d_0^T A(x_0) d_0 + \limsup_{j\to\infty} q_{k_j} d_{k_j}^T b(x_{k_j}) b^T(x_{k_j}) d_{k_j} \le \gamma - \alpha_0. \tag{2.7}$$

Because $k_j \to \infty$ and $q_{k_j} > k_j$ for all j, we must have the following from (2.7):

$$(d_0^T b(x_0))^2 = 0.$$

Thus, (2.7) reduces to $d_0^T A(x_0) d_0 \le \gamma - \alpha_0$ which contradicts (2.3). Thus, (2.5) holds. Combining (2.5) with (2.4) yields $\lim_{q\to+\infty} \mu_q = \gamma$. ∎

We can conclude from Lemma 2.1 that $\gamma > 0$ if and only if there exists a positive number $q_0 > 0$ such that $\mu_q > 0$ when $q > q_0$.

COROLLARY 2.2. (1) *If $\gamma = \lambda$, then for any $q > 0$, $\mu_q = \lambda = \gamma$.*

(2) *If $\gamma > \lambda$, then for any $q > 0$, $\lambda < \mu_q \le \gamma$.*

Proof. (1). It is obvious from Lemma 2.1.

(2). From $\lambda = \min_{x\in X, d\in S^n} d^T A(x)d$, there exist $x_0 \in X$ and $d_0 \in S^n$ such that $\lambda = d_0^T A(x_0)d_0$. Next, we show that $(d_0^T b(x_0))^2 > 0$. If $(d_0^T b(x_0))^2 = 0$, then for any $q > 0$, we have $\mu_q \le q(d_0^T b(x_0))^2 + d_0^T A(x_0)d_0 = \lambda$. Thus, we have $\gamma = \lim_{q\to+\infty} \mu_q \le \lambda$. This contradicts $\gamma > \lambda$. Thus, we must have $(d_0^T b(x_0))^2 > 0$.

From $\mu_q = \min_{x\in X, d\in S^n} d^T \Big(A(x)+qb(x)b^T(x)\Big)d$, there exist $x_q \in X$ and $d_q \in S^n$, such that $\mu_q = q(d_q^T b(x_q))^2 + d_q^T A(x_q)d_q$. Obviously, we have

$$d_q^T A(x_q)d_q \ge d_0^T A(x_0)d_0.$$

If $d_q^T A(x_q)d_q = d_0^T A(x_0)d_0$, then, by the above proof, we have $(d_q^T b(x_q))^2 > 0$. Thus, $\mu_q > d_q^T A(x_q)d_q \ge \lambda$. If $d_q^T A(x_q)d_q > d_0^T A(x_0)d_0$, then we must have $\mu_q > \lambda$. In conclusion, if $\gamma > \lambda$, then for any $q > 0$, we have $\lambda < \mu_q \le \gamma$. ∎

2.2. Positiveness of constrained minimum eigenvalue. Compared to the minimum eigenvalue λ that is defined in (2.2), the exact value of the constrained minimum eigenvalue γ that is defined in (2.3) could be much more difficult to calculate. In certain cases, however, we only need to verify whether $\gamma > 0$ or not. Proposition 5 in [15] suggested a method to verify whether $\gamma > 0$ by working on an $(n+1)\times(n+1)$-bordered Hessian. In the following, we will give another method of verifying whether $\gamma > 0$ by checking whether an $(n-1)\times(n-1)$ matrix is positive definite or not.

We partition symmetric matrix A into the following form:

$$A(x) = \begin{pmatrix} A_1(x) & a(x) \\ a^T(x) & A_2(x) \end{pmatrix}_{n\times n},$$

where

$$A_1(x) = \begin{pmatrix} a_{1,1}(x) & \cdots & a_{1,n-1}(x) \\ \cdots & \cdots & \cdots \\ a_{n-1,1}(x) & \cdots & a_{n-1,n-1}(x) \end{pmatrix} = \Big(a_{i,j}(x)\Big)_{(n-1)\times(n-1)},$$
$$a(x) = \Big(a_{1,n}(x), \ldots, a_{n-1,n}(x)\Big)^T, \quad A_2(x) = \Big(a_{n,n}(x)\Big)_{1\times 1}.$$

For vector $b(x) = (b_1(x), \ldots, b_n(x))^T$ defined in (2.1), we assume that there exists an $i \in \{1, \ldots, n\}$, such that for any $x \in X$, $b_i(x) \ne 0$. Without loss of generality, we assume that $b_n(x) \ne 0$ for all $x \in X$. Let

$$c(x) = \left(\frac{b_1(x)}{b_n(x)}, \ldots, \frac{b_{n-1}(x)}{b_n(x)}\right)^T.$$

Define the following $(n-1)\times(n-1)$ matrix:

$$A_0(x) = A_1(x) - c(x)a^T(x) - a(x)c^T(x) + c(x)A_2(x)c^T(x).$$

THEOREM 2.3. *Assume that $a_{i,j}(x)$, $i, j = 1, \ldots, n$, and $b_i(x)$, $i = 1, \ldots, n$, are continuous on X, and $b_n(x) \ne 0$ for all $x \in X$. Then γ, as defined in* (2.3)*, is strictly positive if and only if $A_0(x)$ is positive definite for all $x \in X$.*

Proof. Because for any $x \in X,\ b_n(x) \neq 0$, we have

$$\begin{aligned}(d^T b(x))^2 = 0 &\Leftrightarrow \left(\sum_{i=1}^{n} b_i(x)d_i\right)^2 = 0\\ &\Leftrightarrow \sum_{i=1}^{n} b_i(x)d_i = 0\\ &\Leftrightarrow d_n = -\sum_{i=1}^{n-1}\frac{b_i(x)}{b_n(x)}d_i\\ &\Leftrightarrow d = \left(d_1, d_2, \ldots, d_{n-1}, -\sum_{i=1}^{n-1}\frac{b_i(x)}{b_n(x)}d_i\right)^T,\\ &= \left(d_1, \ldots, d_{n-1}, -\sum_{i=1}^{n-1} c_i(x)d_i\right)^T\\ &\qquad \text{for any } d_i \in R,\ i = 1, \ldots, n-1.\end{aligned}$$

Let $D_{n-1} = (d_1, \ldots, d_{n-1})^T$. Then, if d satisfies $(d^T b(x))^2 = 0$, then we have

$$d = \Big(D_{n-1}^T,\ -D_{n-1}^T c(x)\Big)^T \tag{2.8}$$

and

$$\begin{aligned}d^T A(x) d &= \Big(D_{n-1}^T,\ -D_{n-1}^T c(x)\Big)\begin{pmatrix} A_1(x) & a(x)\\ a^T(x) & A_2(x)\end{pmatrix}\begin{pmatrix} D_{n-1}\\ -c^T(x)D_{n-1}\end{pmatrix}\\ &= D_{n-1}^T\Big(A_1(x) - c(x)a^T(x) - a(x)c^T(x) + c(x)A_2(x)c^T(x)\Big)D_{n-1}\\ &= D_{n-1}^T A_0(x) D_{n-1}.\end{aligned}$$

From (2.8), we have $d \neq 0$ if and only if $D_{n-1} \neq 0$ and $d^T d = D_{n-1}^T\Big(I_{n-1} + c(x)c^T(x)\Big)D_{n-1}$. Thus, we have

$$\begin{aligned}\gamma &= \min_{x\in X, d\in S^n, b^T(x)d=0} d^T A(x) d\\ &= \min_{x\in X, D_{n-1}^T(c(x)c^T(x)+I_{n-1})D_{n-1}=1} D_{n-1}^T A_0(x) D_{n-1}.\end{aligned}$$

Because $c(x)c^T(x) + I_{n-1}$ is a positive definite matrix, there exists a symmetric nonsingular $(n-1)\times(n-1)$ matrix $E(x)$, such that

$$c(x)c^T(x) + I_{n-1} = E^T(x)E(x).$$

Let $\tilde{d}(x) = E(x)D_{n-1}$. Then,

$$D_{n-1}^T\Big(c(x)c^T(x) + I_{n-1}\Big)D_{n-1} = D_{n-1}^T E^T(x)E(x)D_{n-1} = \tilde{d}^T(x)\tilde{d}(x).$$

Thus, we have

$$\gamma = \min_{x\in X, \tilde{d}\in S^{n-1}} \tilde{d}^T (E^{-1}(x))^T A_0(x) E^{-1}(x)\tilde{d}.$$

Therefore, $\gamma > 0$ if and only if for any $x \in X$, $A_0(x)$ is positive definite. ∎

2.3. Linearly constrained minimum eigenvalue of a constant matrix. In this subsection, we further investigate the relationship between λ and γ, when A is a constant symmetric matrix and $b(x) = Ax + b$, $b \in R^n$. When A is a constant symmetric matrix, all of its eigenvalues can be readily obtained, for example, by using MATLAB programs. Let n eigenvalues of A be arranged as

$$\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n.$$

Clearly, we have

$$\lambda = \min_{d \in S^n} d^T A d = \lambda_1.$$

There exists an orthogonal matrix Ω such that $\Omega^T = \Omega^{-1}$ and

$$\Omega^{-1} A \Omega = \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n).$$

The orthogonal matrix Ω can be readily found by using, for example, the eigenvalue decomposition program in MATLAB. In general, we have

$$\gamma = \min_{(x,d) \in \Gamma} d^T A d \ge \lambda,$$

where $\Gamma = \{(x, d) \mid x \in X, d \in S^n, (Ax + b)^T d = 0\}$. If we let $d = \Omega \bar{d}$, then the expression of γ can be simplified to

$$\gamma = \min_{(x,\bar{d}) \in \bar{\Gamma}} \sum_{k=1}^{n} \lambda_k (\bar{d}_k)^2,$$

where $\bar{\Gamma} = \{(x, \bar{d}) \mid x \in X, \bar{d} \in S^n, (Ax + b)^T \Omega \bar{d} = 0\}$. If we further let $x = \Omega \widetilde{x}$, then the expression of γ can be further simplified to

$$\gamma = \min_{(\widetilde{x},\bar{d}) \in \widetilde{\Gamma}} \sum_{k=1}^{n} \lambda_k (\bar{d}_k)^2,$$

where $\widetilde{\Gamma} = \{(\widetilde{x}, \bar{d}) \mid \widetilde{x} \in \widetilde{X}, \bar{d} \in S^n, \sum_{i=1}^{n} (\lambda_i \widetilde{x}_i \bar{d}_i + \bar{b}_i \bar{d}_i) = 0\}$, $\widetilde{X} = \{\widetilde{x} \in R^n \mid L \le \Omega \widetilde{x} \le U\}$, $\bar{b} = \Omega^T b$, $L = (l_1, \ldots, l_n)^T$, and $U = (u_1, \ldots, u_n)^T$.

Let

$$O = \{i \mid \lambda_i = 0, i = 1, \ldots, n\}$$

and $\bar{O} = \{1, \ldots, n\} \setminus O$. Let $\varsigma_i = \frac{\bar{b}_i}{\lambda_i}$ if $i \in \bar{O}$ and $\varsigma_i = 0$ if $i \in O$. Let $\widetilde{x} = \bar{x}$ - ς, where $\varsigma = (\varsigma_1, \ldots, \varsigma_n)^T$. Finally, the expression of γ can be written in the following form:

$$\gamma = \min_{(\bar{x},\bar{d}) \in \hat{\Gamma}} \sum_{k=1}^{n} \lambda_k (\bar{d}_k)^2,$$

where $\hat{\Gamma} = \{(\bar{x}, \bar{d}) \mid \bar{x} \in \bar{X}, \bar{d} \in S^n, \sum_{i \in \bar{O}} \lambda_i \bar{x}_i \bar{d}_i + \sum_{i \in O} \bar{b}_i \bar{d}_i = 0\}$, $\bar{X} = \{\bar{x} \in R^n \mid L \le \Omega \bar{x} - \Omega \varsigma \le U\}$.

PROPOSITION 2.4. (i) *If* $\lambda_1 = \lambda_2$, *then* $\gamma = \lambda$;

(ii) *if* $0 = \lambda_1 < \lambda_2$, *then*

$$\gamma = \frac{(\bar{b}_1)^2}{(\bar{b}_1)^2 + (\lambda_2)^2 \beta_2} \lambda_2,$$

where $\beta_2 = \max_{x\in\bar{X}}(x_2)^2$*;*

(iii) *if* $\lambda_1 < \lambda_2$, $\lambda_1 \neq 0$, *and there exists a point* $\bar{x} \in \bar{X}$ *such that* $\bar{x}_1 = 0$, *then* $\gamma = \lambda$*;*

(iv) *if* $\lambda_1 < \lambda_2 = 0$ *and there does not exist any point* $\bar{x} \in \bar{X}$ *such that* $\bar{x}_1 = 0$, *then*

$$\gamma = \frac{(\bar{b}_2)^2}{(\bar{b}_2)^2 + (\lambda_1)^2\beta_1}\lambda_1,$$

where $\beta_1 = \max_{x\in\bar{X}}(x_1)^2$*; and*

(v) *if* $0 \neq \lambda_1 < \lambda_2 \neq 0$ *and there does not exist any point* $\bar{x} \in \bar{X}$ *such that* $\bar{x}_1 = 0$, *then*

$$\gamma = \lambda_1 + \frac{\lambda_2 - \lambda_1}{1 + (\frac{\lambda_2}{\lambda_1})^2\beta_{12}},$$

where $\beta_{12} = \max_{x\in\bar{X}}(\frac{x_2}{x_1})^2$.

Proof. (i) If $\lambda_1 = \lambda_2 = \lambda$, then, we have

$$\begin{aligned}\gamma &= \min_{\bar{x}\in\bar{X},\bar{d}\in S^n,\sum_{i\in\bar{O}}\lambda_i\bar{x}_i\bar{d}_i+\sum_{i\in O}\bar{b}_i\bar{d}_i=0}\sum_{k=1}^{n}\lambda_k(\bar{d}_k)^2\\ &= \min_{\bar{x}\in\bar{X},\sum_{i\in\bar{O}\cap\{1,2\}}\lambda_i\bar{x}_i\bar{d}_i+\sum_{i\in O\cap\{1,2\}}\bar{b}_i\bar{d}_i=0,\bar{d}_1^2+\bar{d}_2^2=1}\sum_{k=1}^{2}\lambda_k(\bar{d}_k)^2\\ &= \lambda.\end{aligned}$$

(ii) If $0 = \lambda_1 < \lambda_2$, then $1 \notin \bar{O}$ and $2 \notin O$. Thus,

$$\gamma = \min_{\bar{x}\in\bar{X},\sum_{i\in\bar{O}\cap\{2\}}\lambda_i\bar{x}_i\bar{d}_i+\sum_{i\in O\cap\{1\}}\bar{b}_i\bar{d}_i=0,\bar{d}_1^2+\bar{d}_2^2=1}\sum_{k=1}^{2}\lambda_k(\bar{d}_k)^2.$$

It is clear that if $\bar{b}_1 = 0$, then γ achieves its minimum of 0 by taking $\bar{d}_1 = 1$ and $\bar{d}_2 = 0$. Otherwise, if $\bar{b}_1 \neq 0$, then

$$\gamma = \min_{\bar{x}\in\bar{X},\bar{b}_1\bar{d}_1+\lambda_2\bar{x}_2\bar{d}_2=0,\bar{d}_1^2+\bar{d}_2^2=1}\lambda_2(\bar{d}_2)^2.$$

Because $\bar{b}_1\bar{d}_1 + \lambda_2\bar{x}_2\bar{d}_2 = 0$ and $\bar{d}_1^2 + \bar{d}_2^2 = 1$ imply $(\bar{d}_2)^2 = \frac{1}{1+(\frac{\lambda_2}{\bar{b}_1})^2(\bar{x}_2)^2}$, then we have that

$$\begin{aligned}\gamma &= \min_{\bar{x}\in\bar{X},\bar{b}_1\bar{d}_1+\lambda_2\bar{x}_2\bar{d}_2=0,\bar{d}_1^2+\bar{d}_2^2=1}\lambda_2(\bar{d}_2)^2\\ &= \min_{\bar{x}\in\bar{X}}\frac{\lambda_2}{1+(\frac{\lambda_2}{\bar{b}_1})^2(\bar{x}_2)^2}\\ &= \frac{\lambda_2}{1+(\frac{\lambda_2}{\bar{b}_1})^2\cdot\max_{\bar{x}\in\bar{X}}(\bar{x}_2)^2}\\ &= \frac{(\bar{b}_1)^2}{(\bar{b}_1)^2+(\lambda_2)^2\beta_2}\lambda_2.\end{aligned}$$

Thus, item (ii) of Proposition 2.4 is satisfied for both situations of $\bar{b}_1 = 0$ and $\bar{b}_1 \neq 0$.

(iii) If $\lambda_1 \neq 0$, then $1 \notin O$. Further, since there exists a point $\bar{x} \in \bar{X}$, such that $\bar{x}_1 = 0$, we have

$$\begin{aligned}\gamma &= \min_{\bar{x}\in\bar{X},\sum_{i\in\bar{O}\cap\{1,2\}}\lambda_i\bar{x}_i\bar{d}_i+\sum_{i\in O\cap\{2\}}\bar{b}_i\bar{d}_i=0,\bar{d}_1^2+\bar{d}_2^2=1}\sum_{k=1}^{2}\lambda_k(\bar{d}_k)^2\\ &\leq \min_{\bar{x}\in\{x|x\in\bar{X},\ x_1=0\},\sum_{i\in\bar{O}\cap\{2\}}\lambda_i\bar{x}_i\bar{d}_i+\sum_{i\in O\cap\{2\}}\bar{b}_i\bar{d}_i=0,\bar{d}_1^2+\bar{d}_2^2=1}\sum_{k=1}^{2}\lambda_k(\bar{d}_k)^2\\ &\leq \lambda_1(\text{where we take } \bar{d}_1 = 1, \bar{d}_2 = 0).\end{aligned}$$

Because $\gamma \geq \lambda_1$, we have $\gamma = \lambda_1 = \lambda$.

(iv) Because $\lambda_1 < \lambda_2 = 0$, then $1 \notin O$ and $2 \notin \bar{O}$. We have

$$\begin{aligned}\gamma &= \min_{\bar{x}\in\bar{X},\sum_{i\in\bar{O}\cap\{1\}}\lambda_i\bar{x}_i\bar{d}_i+\sum_{i\in O\cap\{2\}}\bar{b}_i\bar{d}_i=0,\bar{d}_1^2+\bar{d}_2^2=1}\sum_{k=1}^{2}\lambda_k(\bar{d}_k)^2\\ &= \min_{\bar{x}\in\bar{X},\lambda_1\bar{x}_1\bar{d}_1+\bar{b}_2\bar{d}_2=0,\bar{d}_1^2+\bar{d}_2^2=1}\lambda_1(\bar{d}_1)^2.\end{aligned}$$

If $\bar{b}_2 = 0$, then we must have $\bar{d}_1 = 0$ because $\lambda_1 < 0$, and there does not exist any $\bar{x} \in \bar{X}$, such that $\bar{x}_1 = 0$. Thus, $\gamma = 0$ when $\bar{b}_2 = 0$. Otherwise, when $\bar{b}_2 \neq 0$, $\lambda_1\bar{x}_1\bar{d}_1 + \bar{b}_2\bar{d}_2 = 0$ and $\bar{d}_1^2 + \bar{d}_2^2 = 1$ imply $(\bar{d}_1)^2 = \frac{1}{1+(\frac{\lambda_1}{b_2})^2\bar{x}_1^2}$. We then have

$$\begin{aligned}\gamma &= \min_{\bar{x}\in\bar{X},\lambda_1\bar{x}_1\bar{d}_1+\bar{b}_2\bar{d}_2=0,\bar{d}_1^2+\bar{d}_2^2=1}\lambda_1(\bar{d}_1)^2\\ &= \frac{\lambda_1}{1+(\frac{\lambda_1}{\bar{b}_2})^2\max\limits_{\bar{x}\in\bar{X}}(\bar{x}_1)^2}\\ &= \frac{(\bar{b}_2)^2}{(\bar{b}_2)^2+(\lambda_1)^2\beta_1}\lambda_1.\end{aligned}$$

Thus, item (iv) of Proposition 2.4 is satisfied for both situations of $\bar{b}_2 = 0$ and $\bar{b}_2 \neq 0$.

(v) Because both λ_1 and λ_2 are not equal to zero, neither 1 nor 2 belongs to O. Then,

$$\gamma = \min_{\bar{x}\in\bar{X},\lambda_1\bar{x}_1\bar{d}_1+\lambda_2\bar{x}_2\bar{d}_2=0,\bar{d}_1^2+\bar{d}_2^2=1}\sum_{k=1}^{2}\lambda_k(\bar{d}_k)^2.$$

Because for any $\bar{x} \in \bar{X}$, $\bar{x}_1 \neq 0$, then $\lambda_1\bar{x}_1\bar{d}_1 + \lambda_2\bar{x}_2\bar{d}_2 = 0$ and $\bar{d}_1^2 + \bar{d}_2^2 = 1$ imply that $(\bar{d}_2)^2 = \frac{1}{1+(\frac{\lambda_2}{\lambda_1})^2(\frac{\bar{x}_2}{\bar{x}_1})^2}$. Thus, we have

$$\begin{aligned}\gamma &= \min_{\bar{x}\in\bar{X},\lambda_1\bar{x}_1\bar{d}_1+\lambda_2\bar{x}_2\bar{d}_2=0,\bar{d}_1^2+\bar{d}_2^2=1}\sum_{k=1}^{2}\lambda_k(\bar{d}_k)^2\\ &= \min_{\bar{x}\in\bar{X}}\left[\left(1-\frac{1}{1+(\frac{\lambda_2}{\lambda_1})^2(\frac{\bar{x}_2}{\bar{x}_1})^2}\right)\lambda_1+\frac{1}{1+(\frac{\lambda_2}{\lambda_1})^2(\frac{\bar{x}_2}{\bar{x}_1})^2}\lambda_2\right]\\ &= \lambda_1+\frac{\lambda_2-\lambda_1}{1+(\frac{\lambda_2}{\lambda_1})^2\max_{\bar{x}\in\bar{X}}(\frac{\bar{x}_2}{\bar{x}_1})^2}\\ &= \lambda_1+\frac{\lambda_2-\lambda_1}{1+(\frac{\lambda_2}{\lambda_1})^2\beta_{12}}. \qquad \square\end{aligned}$$

Remark 2.1. (i) To judge whether there exists a point $\bar{x} \in \bar{X}$ such that $\bar{x}_1 = 0$, two linear programming problems can be constructed, $\{\max x_1 \mid x \in \bar{X}\}$ and $\{\min x_1 \mid x \in \bar{X}\}$. If $\max_{x\in\bar{X}} x_1 < 0$ or $\min_{x\in\bar{X}} x_1 > 0$, then there does not exist an $\bar{x} \in \bar{X}$, such that $\bar{x}_1 = 0$. Otherwise, there exists one.

(ii) For β_1 in (iv) and β_2 in (ii) of the above proposition, the convex maximization (or, equivalently, concave minimization) problem of $\{\max(x_i)^2 \mid x \in \bar{X}\}$, i = 1, 2, can be replaced by two linear programming problems, $\{\max x_i \mid x \in \bar{X}\}$ and $\{\min x_i \mid x \in \bar{X}\}$, because $\max_{x\in\bar{X}}(x_i)^2 = \max\{(\max_{x\in\bar{X}} x_i)^2, (\min_{x\in\bar{X}} x_i)^2\}$.

(iii) Consider β_{12} in (v) of the above proposition. Let $y_1 = \frac{1}{x_1}$ and $y_i = \frac{x_i}{x_1}$, $i = 2, \ldots, n$. If for every $\bar{x} \in \bar{X}$, $\bar{x}_1 > 0$, then problem $\{\max(\frac{x_2}{x_1})^2 \mid x \in \bar{X}\}$ is equivalent to $\{\max(y_2)^2 \mid y \in \bar{Y}_1\}$, where

$$\bar{Y}_1 = \{y \in R^n \mid y_1 L \le \Omega(1, y_2, \ldots, y_n)^T - \Omega\varsigma y_1 \le y_1 U\},$$

whereas convex maximization (or, equivalently, concave minimization) problem $\{\max(y_2)^2 \mid y \in \bar{Y}_1\}$ can be replaced by two linear programming problems, $\{\max y_2 \mid y \in \bar{Y}_1)$ and $(\min y_2 \mid y \in \bar{Y}_1\}$, because $\max_{y\in\bar{Y}_1}(y_2)^2 = \max\{\big(\max_{y\in\bar{Y}_1} y_2\big)^2, \big(\min_{y\in\bar{Y}_1} y_2\big)^2\}$. Similarly, if for every $\bar{x} \in \bar{X}$, $\bar{x}_1 < 0$, then $\max_{x\in\bar{X}}(\frac{x_2}{x_1})^2 = \max\{\big(\max_{y\in\bar{Y}_2} y_2\big)^2, \big(\min_{y\in\bar{Y}_2} y_2\big)^2\}$, where

$$\bar{Y}_2 = \{y \in R^n \mid y_1 L \ge \Omega(1, y_2, \ldots, y_n)^T - \Omega\varsigma y_1 \ge y_1 U\}.$$

Based on the above proposition, the computation of the linearly constrained minimum eigenvalue of a constant matrix can be carried out by the following algorithm.

Algorithm I (Calculation of γ).

Step 1. Calculate an orthogonal matrix Ω such that

$$\Omega^T A \Omega = \operatorname{diag}(\lambda_1, \ldots, \lambda_n),$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A satisfying $\lambda_1 \le \cdots \le \lambda_n$. Let $\bar{b} = (\bar{b}_1, \ldots, \bar{b}_n)^T := \Omega^T b$ and $\varsigma = (\varsigma_1, \ldots, \varsigma_n)^T$, where $\varsigma_i := \begin{cases} \frac{\bar{b}_i}{\lambda_i} & \text{if } \lambda_i \ne 0, \\ 0 & \text{if } \lambda_i = 0. \end{cases}$

Step 2. If $\lambda_1 = \lambda_2$, then $\gamma = \lambda_1$ and stop; otherwise go to Step 3.

Step 3. If $\lambda_1 = 0$, then solve the following linear programming problems:

$$(P_1) \qquad \begin{aligned} &\max x_2 \\ &L \le \Omega(x - \varsigma) \le U \end{aligned}$$

and

$$(P_2) \qquad \begin{aligned} &\min x_2 \\ &L \le \Omega(x - \varsigma) \le U. \end{aligned}$$

Let v_1 and v_2 be the optimal values of problems (P_1) and (P_2), respectively. Let $\beta_2 = \max\{v_1^2, v_2^2\}$. Then $\gamma = \frac{(\bar{b}_1)^2\lambda_2}{(\bar{b}_1)^2 + (\lambda_2)^2\beta_2}$ and stop. If $\lambda_1 \ne 0$, then go to Step 4.

Step 4. Solve the following linear programming problems:

$$(P_3) \qquad \begin{aligned} &\max x_1 \\ &L \le \Omega(x - \varsigma) \le U \end{aligned}$$

and

$$(P_4) \qquad \begin{aligned} &\min x_1 \\ &L \le \Omega(x-\varsigma) \le U. \end{aligned}$$

Let v_3 and v_4 be the optimal values of problems (P_3) and (P_4), respectively. If $v_3 \ge 0 \ge v_4$, then $\gamma = \lambda_1$ and stop; otherwise go to Step 5.

Step 5. If $\lambda_2 = 0$, let $\beta_1 = \max\{v_3^2, v_4^2\}$, then $\gamma = \frac{(\bar{b}_2)^2\lambda_1}{(\bar{b}_2)^2+(\lambda_1)^2\beta_1}$ and stop; otherwise go to Step 6.

Step 6. If $v_4 > 0$, then solve the following linear programming problems:

$$(P_5) \qquad \begin{aligned} &\max y_2 \\ &y_1 L \le \Omega(1-y_1\varsigma_1, y_2-y_1\varsigma_2, \ldots, y_n-y_1\varsigma_n)^T \le y_1 U \end{aligned}$$

and

$$(P_6) \qquad \begin{aligned} &\min y_2 \\ &y_1 L \le \Omega(1-y_1\varsigma_1, y_2-y_1\varsigma_2, \ldots, y_n-y_1\varsigma_n)^T \le y_1 U. \end{aligned}$$

Let v_5 and v_6 be the optimal values of problems (P_5) and (P_6), respectively. Let $\beta_3 = \max\{v_5^2, v_6^2\}$. Then, $\gamma = \lambda_1 + \frac{\lambda_2-\lambda_1}{1+\left(\frac{\lambda_2}{\lambda_1}\right)^2\beta_3}$ and stop; otherwise, solve the following linear programming problems:

$$(P_7) \qquad \begin{aligned} &\max y_2 \\ &y_1 L \ge \Omega(1-y_1\varsigma_1, y_2-y_1\varsigma_2, \ldots, y_n-y_1\varsigma_n)^T \ge y_1 U \end{aligned}$$

and

$$(P_8) \qquad \begin{aligned} &\min y_2 \\ &y_1 L \ge \Omega(1-y_1\varsigma_1, y_2-y_1\varsigma_2, \ldots, y_n-y_1\varsigma_n)^T \ge y_1 U. \end{aligned}$$

Let v_7 and v_8 be the optimal values of problems (P_7) and (P_8), respectively. Let $\beta_4 = \max\{v_7^2, v_8^2\}$. Then $\gamma = \lambda_1 + \frac{\lambda_2-\lambda_1}{1+\left(\frac{\lambda_2}{\lambda_1}\right)^2\beta_4}$ and stop.

3. Hidden-convex function.

DEFINITION 3.1 (see [11]). *A function $f : R^n \to R$ is strictly increasing (decreasing) on X with respect to x_i if*

$$f(x_1, \ldots, x_{i-1}, \tilde{x}_i, x_{i+1}, \ldots, x_n) < (>)\ f(x_1, \ldots, x_{i-1}, \bar{x}_i, x_{i+1}, \ldots, x_n)$$

for any $(x_1, \ldots, x_{i-1}, \tilde{x}_i, x_{i+1}, \ldots, x_n)^T$, $(x_1, \ldots, x_{i-1}, \bar{x}_i, x_{i+1}, \ldots, x_n)^T \in X$, and $\tilde{x}_i < (>)\ \bar{x}_i$.

DEFINITION 3.2. *A function f: $R^n \to R$ is said to be strictly monotone if for any $i = 1, \ldots, n$, f is either strictly increasing or strictly decreasing on X with respect to x_i.*

Let function h be a second order continuously differentiable function that is defined on the box X given in (1.1). Consider the following three transformations of function h:

$$\begin{aligned} \phi(y) &= h(t(y)), \\ \varphi(x) &= T(h(x)), \\ \psi(y) &= T(h(t(y))), \end{aligned} \tag{3.1}$$

where $T : R^1 \to R^1$ is a real function and $t : R^n \to R^n$ is a separable mapping, i.e., $t(y) = \Big(t_1(y_1), \ldots, t_n(y_n)\Big)^T$ for $y = (y_1, \ldots, y_n)^T$. We further assume that each t_i, $i = 1, \cdots, n$, is a strictly monotone mapping. It is clear that the domain of ϕ and ψ is

$$\begin{aligned} Y &= \{y \in R^n \mid y_i = t_i^{-1}(x_i), i = 1, \ldots, n,\ x \in X\} \\ &= \left\{ y \in R^n \;\middle|\; \begin{array}{ll} t_i^{-1}(l_i) \le y_i \le t_i^{-1}(u_i), & t_i(y_i) \text{ is strictly increasing,} \\ t_i^{-1}(u_i) \le y_i \le t_i^{-1}(l_i), & t_i(y_i) \text{ is strictly decreasing} \end{array} \right\}. \end{aligned}$$

Obviously, Y is also a box. For the purpose of convenience, let $\min_{x\in\emptyset} x = +\infty$ and $\max_{x\in\emptyset} x = -\infty$, where $x \in R$.

DEFINITION 3.3. *If there exists a separable strictly monotone mapping $x = t(y)$, such that $\phi(y) = h(t(y))$ is a (strictly) convex function on Y, then h is called a* d-hidden (strictly) convex function *on X.*

DEFINITION 3.4. *If there exists a strictly increasing and second order continuously differentiable function $T(\cdot) : h(X) \to R$ satisfying $T'(y) > 0$ for any $y \in h(X)$, such that $\varphi(x) = T(h(x))$ is a (strictly) convex function on X, then h is called an* r-hidden (strictly) convex function *on X.*

The r-hidden convex function is also called a G-convex function on X [4].

DEFINITION 3.5. *If there exist a strictly increasing and second order continuously differentiable function $T(\cdot)$ satisfying $T'(y) > 0$ for any $y \in h(X)$ and a separable strictly monotone mapping $x = t(y)$, such that the function $\psi(y) = T(h(t(y)))$ is (strictly) convex on Y, then h is called a* hidden (strictly) convex function *on X.*

Obviously, a convex function is also d-hidden convex, r-hidden convex, and hidden convex. From Definitions 3.3, 3.4, and 3.5, we know that if function h is d-hidden convex or r-hidden convex, then h is also hidden convex. From [1], we know that if function h is r-hidden convex, then h must be pseudoconvex. Because the hidden-convex function that is defined in Definition 3.5 confines its variable transformation to being separable, the hidden-convex function is thus a special case of the (t, T)-convex function that is defined in [8]. It is well known that any local minimizer of a pseudoconvex function on a convex set is also a global minimizer. In this paper, we will also show that any local minimizer of a hidden-convex function on a box set is also a global minimizer.

3.1. Necessary and sufficient conditions for r-hidden convex functions. Notice that

$$\begin{aligned} \frac{\partial\varphi(x)}{\partial x_i} &= T'(h(x))\frac{\partial h(x)}{\partial x_i},\ i = 1, \ldots, n, \\ \frac{\partial^2\varphi(x)}{\partial x_i^2} &= T'(h(x))\left[\frac{T''(h(x))}{T'(h(x))}\left(\frac{\partial h(x)}{\partial x_i}\right)^2 + \frac{\partial^2 h(x)}{\partial x_i^2}\right],\ i = 1, \ldots, n, \\ \frac{\partial^2\varphi(x)}{\partial x_i \partial x_j} &= T'(h(x))\left[\frac{T''(h(x))}{T'(h(x))}\frac{\partial h(x)}{\partial x_i}\frac{\partial h(x)}{\partial x_j} + \frac{\partial^2 h(x)}{\partial x_i \partial x_j}\right], \\ & \qquad i, j = 1, \ldots, n,\ i \ne j. \end{aligned}$$

Let $H(x)$ and $\bar{H}(x)$ be the Hessian of h and φ at x, respectively. Also let $a_h(x) = \nabla h(x) = \left(\frac{\partial h(x)}{\partial x_1}, \ldots, \frac{\partial h(x)}{\partial x_n}\right)^T$ be the gradient of $h(x)$ at x. Thus, we have

$$\bar{H}(x) = T'(h(x))\left[\frac{T''(h(x))}{T'(h(x))}a_h(x)a_h^T(x) + H(x)\right].$$

Let

$$\begin{aligned}\lambda^h &= \min_{x\in X, d\in S^n} d^T H(x) d,\\ \gamma^h &= \min_{x\in X, d\in S^n, a_h^T(x)d=0} d^T H(x) d,\\ B_q^h(x) &= q a_h(x) a_h^T(x) + H(x),\\ \mu_q^h &= \min_{x\in X, d\in S^n} d^T B_q^h(x) d.\end{aligned}$$

THEOREM 3.6. *Given a second order differentiable function h defined on the box set X, the following statements hold:*

(i) *h is an r-hidden strictly convex function if $\gamma^h > 0$ and*

(ii) *$\gamma^h \geq 0$ if h is an r-hidden convex function.*

Proof. (i) Clearly, h is an r-hidden strictly convex function if for any $x \in X$, $\bar{H}(x)$ is positive definite. Furthermore, because $T'(h(x)) > 0$ for any $x \in X$, $\bar{H}(x)$ is positive definite if and only if $\dfrac{T''(h(x))}{T'(h(x))} a_h(x) a_h^T(x) + H(x)$ is positive definite.

If we take $T(s) = \exp(qs)$ with $q > 0$, then T satisfies the following conditions:

$$T'(s) > 0, \ \frac{T''(s)}{T'(s)} = q \quad \forall s \in R.$$

Thus,

$$\frac{T''(h(x))}{T'(h(x))} a_h(x) a_h^T(x) + H(x) = q a_h(x) a_h^T(x) + H(x).$$

By Lemma 2.1, if $\gamma^h > 0$, then there exists a positive number $q > 0$, such that $\mu_q^h > 0$, i.e., $q a_h(x) a_h^T(x) + H(x)$ is positive definite on X. Thus, $\varphi(x)$ is strictly convex when $T(s)$ is selected as $\exp(qs)$ with a large enough q. This implies that h is an r-hidden strictly convex function on X.

(ii) If $h(x)$ is an r-hidden convex function, then there must exist a second order continuously differentiable function T, satisfying $T'(h(x)) > 0$ for any $x \in X$, such that $\dfrac{T''(h(x))}{T'(h(x))} a_h(x) a_h^T(x) + H(x)$ is a positive semidefinite matrix, i.e.,

$$\eta = \min_{x\in X, \ d\in S^n} d^T \left(\frac{T''(h(x))}{T'(h(x))} a_h(x) a_h^T(x) + H(x) \right) d \geq 0.$$

Because

$$\begin{aligned}\gamma^h &= \min_{x\in X, d\in S^n, d^T a_h(x)=0} d^T H(x) d\\ &= \min_{x\in X, d\in S^n, d^T a_h(x)=0} d^T \left(\frac{T''(h(x))}{T'(h(x))} a_h(x) a_h^T(x) + H(x) \right) d\\ &\geq \eta.\end{aligned}$$

Thus, we have that $\gamma^h \geq 0$. ∎

Theorem 3.6 gives a sufficient condition under which a function h can be converted into a strictly convex function only by a range transformation and a necessary condition under which a function h can be converted into a convex function only by

a range transformation. As is well known, these r-hidden convex functions must be pseudoconvex. Notice that a sufficient condition for pseudoconvexity of a second order differentiable function [5] is that for all $x \in X$, $\frac{1}{2}[\delta - h(x)]a_h(x)a_h^T(x) + H(x)$ is positive semidefinite for some $\delta > h(x)$.

In the following, we will give a condition under which a function can be converted into a convex function using both a range transformation and a domain transformation.

3.2. Sufficient conditions for hidden-convex functions. Let index set $NZ \subset \{1, \ldots, n\}$ be defined as

$$NZ = \left\{ i \in \{1, \ldots, n\} \mid \frac{\partial h(x)}{\partial x_i} \neq 0 \ \text{ for any } x \in X \right\}. \tag{3.2}$$

If $NZ \neq \emptyset$, then we assume, without loss of generality, that $NZ = \{l+1, \ldots, n\}$, where $l \in \{0, 1, \ldots, n-1\}$; otherwise, we can appropriately rearrange x_i, $i = 1, \ldots, n$. If $NZ = \emptyset$, then we let $l = n$. Note that l is the number of indices such that whenever $l > 0$ there exists an x, such that $\frac{\partial h(x)}{\partial x_i} = 0$ for $1 \leq i \leq l$.

For a given second order differentiable function h, let

$$H_k(x) = \begin{pmatrix} \dfrac{\partial^2 h(x)}{\partial x_1^2} & \cdots & \dfrac{\partial^2 h(x)}{\partial x_1 \partial x_k} \\ \vdots & \vdots & \vdots \\ \dfrac{\partial^2 h(x)}{\partial x_k \partial x_1} & \cdots & \dfrac{\partial^2 h(x)}{\partial x_k^2} \end{pmatrix}, \ k = 1, \ldots, n, \tag{3.3}$$

$$a_{h,k}(x) = \left(\frac{\partial h(x)}{\partial x_1}, \ldots, \frac{\partial h(x)}{\partial x_k} \right)^T, \ k = 1, \ldots, n, \tag{3.4}$$

$$\lambda_k^h = \min_{x \in X, d \in S^k} d^T (H_k(x)) d, \ k = 1, \ldots, n,$$

$$\gamma_k^h = \min_{x \in X, d \in S^k, a_{h,k}^T(x) d = 0} d^T (H_k(x)) d, \ k = 1, \ldots, n, \tag{3.5}$$

$$\Xi_k(x) = \left(\frac{\partial^2 h(x)}{\partial x_1 \partial x_k}, \ldots, \frac{\partial^2 h(x)}{\partial x_{k-1} \partial x_k} \right)^T, \ k = 2, \ldots, n,$$

$$C_{p,k}(x) = \operatorname{diag}\left(p_1 \frac{\partial h(x)}{\partial x_1}, \ldots, p_k \frac{\partial h(x)}{\partial x_k} \right), \ k = 1, \ldots, n, \tag{3.6}$$

$$\text{where } p_i \in R, \ i = 1, \ldots, k,$$

$$\nu_{k,p}^h = \min_{x \in X, d \in S^k, a_{h,k}^T(x) d = 0} d^T (H_k(x) + C_{p,k}(x)) d, \ k = 1, \ldots, n,$$

$$\nu_k^h = \sup_{p \in R^k} \nu_{k,p}^h.$$

If $l < n$, then let

$$\xi_k(x) = \left(\frac{\partial h(x)}{\partial x_1} / \frac{\partial h(x)}{\partial x_{k+1}}, \ldots, \frac{\partial h(x)}{\partial x_k} / \frac{\partial h(x)}{\partial x_{k+1}} \right)^T, \ k = l, \ldots, n-1.$$

Notice $H_n(x) = H(x)$, $a_{h,n}(x) = a_h(x)$, $\lambda_n^h = \lambda^h$, and $\gamma_n^h = \gamma^h$.

Furthermore, define the following set:

$$P_k = \{ p = (p_1, \ldots, p_k)^T \mid p \text{ is chosen such that } \nu_{k,p}^h > 0 \}, \ k = 1, \ldots, n.$$

For given $p_1, \ldots, p_n \in R$, let

$$t_{i,p_i}(y_i) = \begin{cases} y_i & p_i = 0, \\ -\dfrac{1}{p_i} \ln y_i & p_i \neq 0, \end{cases} \quad i = 1, \ldots, n,$$

$$Y_p = \{y \in R^n \mid (t_{1,p_1}(y_1), \ldots, t_{n,p_n}(y_n))^T \in X\},$$

$$Y_{i,p} = \{y_i \mid \exists y_j,\ j = 1, \ldots, i-1, i+1, \ldots, n \text{ such that } (y_1, \ldots, y_{i-1}, y_i, y_{i+1}, \ldots, y_n)^T \in Y_p\}.$$

Then, $t_{i,p_i}(y_i)$ is strictly monotone on $Y_{i,p}$ and satisfies

$$t'_{i,p_i}(y_i) \neq 0,$$
$$\frac{t''_{i,p_i}(y_i)}{\left[t'_{i,p_i}(y_i)\right]^2} = p_i \ \text{ for any } y_i \in Y_{i,p},\ i = 1, \ldots, n.$$

For given $q > 0$, let $T_q(s) = \exp(qs)$. Then $T_q(s)$ satisfies

$$T'_q(s) > 0,\ \frac{T''_q(s)}{T'_q(s)} = q \quad \forall s \in R.$$

Let $x = t_p(y) = (t_{1,p_1}(y_1), \ldots, t_{n,p_n}(y_n))^T$ and $\psi_{q,p}(y) = T_q(h(t_p(y)))$ for all $y \in Y_p$. Then, we have

$$\begin{aligned}
\frac{\partial \psi_{q,p}(y)}{\partial y_k} &= T'_q(h(x)) \frac{\partial h(x)}{\partial x_k} t'_{k,p_k}(y_k), \\
\frac{\partial^2 \psi_{q,p}(y)}{\partial y_k^2} &= T''_q(h(x)) \left[\frac{\partial h(x)}{\partial x_k} t'_{k,p_k}(y_k)\right]^2 + T'_q(h(x)) \frac{\partial^2 h(x)}{\partial x_k^2} [t'_{k,p_k}(y_k)]^2 \\
&\quad + T'_q(h(x)) \frac{\partial h(x)}{\partial x_k} t''_{k,p_k}(y_k) \\
&= T'_q(h(x)) [t'_{k,p_k}(y_k)]^2 \\
&\quad \times \left[\frac{T''_q(h(x))}{T'_q(h(x))} \left(\frac{\partial h(x)}{\partial x_k}\right)^2 + \frac{\partial^2 h(x)}{\partial x_k^2} + \frac{\partial h(x)}{\partial x_k} \frac{t''_{k,p_k}(y_k)}{[t'_{k,p_k}(y_k)]^2}\right] \\
&= T'_q(h(x)) [t'_{k,p_k}(y_k)]^2 \left[q\left(\frac{\partial h(x)}{\partial x_k}\right)^2 + \frac{\partial^2 h(x)}{\partial x_k^2} + p_k \frac{\partial h(x)}{\partial x_k}\right], \\
\frac{\partial^2 \psi_{q,p}(y)}{\partial y_k \partial y_j} &= T'_q(h(x)) t'_{k,p_k}(y_k) t'_{j,p_j}(y_j) \left[q \frac{\partial h(x)}{\partial x_k} \frac{\partial h(x)}{\partial x_j} + \frac{\partial^2 h(x)}{\partial x_k \partial x_j}\right], \\
&\qquad \text{for } k \neq j.
\end{aligned}$$

THEOREM 3.7. *Given a second order continuously differentiable function h, if $l \geq 1$ and $\nu_l^h > 0$, i.e., $P_l \neq \emptyset$, then h is a hidden strictly convex function.*

Proof. Let $\psi(y)$ in (3.1) be $\psi_{q,p}(y)$. Let $H_{q,p}(y)$ be Hessian of $\psi_{q,p}$ at y. Then, we have

$$H_{q,p}(y) = T'_q(h(x)) S(x) [q a_{h,n}(x) a_{h,n}^T(x) + H_n(x) + C_{p,n}(x)] S(x),$$

where

$$S(x) = \operatorname{diag}\Big(t'_{1,p_1}(y_1), t'_{2,p_2}(y_2), \ldots, t'_{n,p_n}(y_n)\Big),$$

and $H_n(x)$, $a_{h,n}(x)$ and $C_{p,n}(x)$ are defined in (3.3), (3.4), and (3.6), respectively.

Obviously, $H_{q,p}(y)$ is positive definite if and only if $qa_{h,n}(x)a_{h,n}^T(x) + H_n(x) + C_{p,n}(x)$ is positive definite. By Lemma 2.1, for a given vector $p \in R^n$, there exists a positive number q_0 such that $q_0 a_{h,n}(x)a_{h,n}^T(x) + H_n(x) + C_{p,n}(x)$ is positive definite if and only if $\nu_{n,p}^h > 0$. Note that

$$H_n(x) + C_{p,n}(x) = \begin{pmatrix} H_{n-1}(x) + C_{p,n-1}(x) & \Xi_n(x) \\ \Xi_n^T(x) & \dfrac{\partial^2 h(x)}{\partial x_n^2} + p_n \dfrac{\partial h(x)}{\partial x_n} \end{pmatrix}.$$

If $n \in NZ$, then, by Theorem 2.3, $\nu_{n,p}^h > 0$ if and only if $H_{n-1}(x) + C_{p,n-1}(x) - \Xi_n(x)\xi_{n-1}^T(x) - \xi_{n-1}(x)\Xi_n^T(x) + \left(\dfrac{\partial^2 h(x)}{\partial x_n^2} + p_n \dfrac{\partial h(x)}{\partial x_n}\right)\xi_{n-1}(x)\xi_{n-1}^T(x)$ is positive definite. Note that

$$\begin{aligned} &H_{n-1}(x) + C_{p,n-1}(x) - \Xi_n(x)\xi_{n-1}^T(x) - \xi_{n-1}(x)\Xi_n^T(x) \\ &+ \left(\frac{\partial^2 h(x)}{\partial x_n^2} + p_n \frac{\partial h(x)}{\partial x_n}\right)\xi_{n-1}(x)\xi_{n-1}^T(x) \\ &= H_{n-1}(x) + C_{p,n-1}(x) + (\xi_{n-1}(x) - \Xi_n(x))(\xi_{n-1}(x) - \Xi_n(x))^T \\ &- \Xi_n(x)\Xi_n^T(x) + \left(\frac{\partial^2 h(x)}{\partial x_n^2} + p_n \frac{\partial h(x)}{\partial x_n} - 1\right)\xi_{n-1}(x)\xi_{n-1}^T(x). \end{aligned}$$

For any $k \in NZ$, $\dfrac{\partial h(x)}{\partial x_k} \neq 0$ for any $x \in X$, i.e., $\left|\dfrac{\partial h(x)}{\partial x_k}\right| > 0$ for any $x \in X$. Let

$$\begin{aligned} m_k &= \min_{x \in X} \left|\frac{\partial h(x)}{\partial x_k}\right|, \\ \alpha_k &= \min_{x \in X} \frac{\partial^2 h(x)}{\partial x_k^2}, \\ q_k &= |p_k| m_k + \alpha_k - 1, \\ W_{k,p}(x) &= H_k(x) + C_{p,k}(x) + (\xi_k(x) - \Xi_{k+1}(x))(\xi_k(x) - \Xi_{k+1}(x))^T \\ &\quad -\Xi_{k+1}(x)\Xi_{k+1}^T(x). \end{aligned} \tag{3.7}$$

For any $k \in NZ$, let p_k satisfy

$$\operatorname{sign}(p_k) = \operatorname{sign}\left(\frac{\partial h(x)}{\partial x_k}\right), \tag{3.8}$$

where $\operatorname{sign}(t) = \begin{cases} 1 & t > 0, \\ 0 & t = 0, \\ -1 & t < 0. \end{cases}$

Then, for any $x \in X$, we have $\dfrac{\partial^2 h(x)}{\partial x_k^2} + p_k \dfrac{\partial h(x)}{\partial x_k} - 1 \geq q_k$.

If q_n and p_n are selected according to (3.7) and (3.8), respectively, then for any $x \in X$: if $W_{n-1,p}(x) + q_n \xi_{n-1}(x)\xi_{n-1}^T(x)$ is a positive definite matrix, then $W_{n-1,p}(x) + \left(\dfrac{\partial^2 h(x)}{\partial x_n^2} + p_n \dfrac{\partial h(x)}{\partial x_n} - 1\right)\xi_{n-1}(x)\xi_{n-1}^T(x)$ must be positive definite.

By Lemma 2.1, there exists a $q_n > 0$, such that for any $x \in X$, $W_{n-1,p}(x) + q_n \xi_{n-1}(x)\xi_{n-1}^T(x)$ is positive definite if and only if

$$\min_{x\in X, d\in S^{n-1}, d^T\xi_{n-1}(x)=0} d^T W_{n-1,p}(x) d > 0.$$

Note that

$$\begin{aligned}
&\min_{x\in X, d\in S^{n-1}, d^T\xi_{n-1}(x)=0} d^T W_{n-1,p}(x) d \\
&= \min_{x\in X, d\in S^{n-1}, d^T\xi_{n-1}(x)=0} \Big[d^T \Big(H_{n-1}(x) + C_{p,n-1}(x) \Big) d - d^T \Xi_n(x) \Xi_n^T(x) d \\
&\quad + d^T \Big((\xi_{n-1}(x) - \Xi_n(x))(\xi_{n-1}(x) - \Xi_n(x))^T \Big) d \Big] \\
&= \min_{x\in X, d\in S^{n-1}, d^T a_{h,n-1}(x)=0} d^T \Big(H_{n-1}(x) + C_{p,n-1}(x) \Big) d \\
&= \nu_{n-1,p}^h.
\end{aligned}$$

Thus, if $\nu_{n-1,p}^h > 0$, then there exists a $q_n^* > 0$, such that for any $x \in X$, $W_{n-1,p}(x) + q_n^* \xi_{n-1}(x)\xi_{n-1}^T(x)$ is positive definite. If we take p_n satisfying (3.8), and $|p_n|$ is large enough such that $q_n = |p_n| m_n + \alpha_n - 1 \geq q_n^*$, then for any $x \in X$, $W_{n-1,p}(x) + \left(\dfrac{\partial^2 h(x)}{\partial x_n^2} + p_n \dfrac{\partial h(x)}{\partial x_n} - 1 \right) \xi_{n-1}(x)\xi_{n-1}^T(x)$ is positive definite. Because for any $x \in X$, $W_{n-1,p}(x) + \left(\dfrac{\partial^2 h(x)}{\partial x_n^2} + p_n \dfrac{\partial h(x)}{\partial x_n} - 1 \right) \xi_{n-1}(x)\xi_{n-1}^T(x)$ is positive definite if and only if $\nu_{n,p}^h > 0$. Thus, we must have $\nu_n^h > 0$.

Similarly, we can prove that if $n-1 \in NZ$ and $\nu_{n-2}^h > 0$, then we can properly choose p_{n-1} such that $\nu_{n-1}^h > 0$. Repeating this process, we can prove for all $k = n-3$, $n-4, \ldots, l$, that if $\nu_k^h > 0$, then we can properly choose p_{k+1} such that $\nu_{k+1}^h > 0$.

From the above discussion, we know that if $l \geq 1$ and $\nu_l^h > 0$, i.e., $P_l \neq \emptyset$, then we can properly choose p_k, $k = l+1, \ldots, n$, such that $\nu_n^h > 0$. Note that $\nu_n^h > 0$ if and only if there exist a positive number $q_0 > 0$ and a vector $p \in R^n$, such that for any $x \in X$, $q_0 a_{h,n}(x) a_{h,n}^T(x) + H_n(x) + C_{p,n}(x)$ is positive definite. Thus, $H_{q_0,p}(y)$ is positive definite on Y_p which implies that $\psi_{q_0,p} = T_{q_0}(h(t_p(y)))$ is strictly convex on Y_p. Thus, if $l \geq 1$ and $\nu_l^h > 0$, i.e., $P_l \neq \emptyset$, then $h(x)$ must be a hidden strictly convex function on X. □

Based on Theorem 3.7, we have the following corollary.

COROLLARY 3.8. *If a second order continuously differentiable function h defined on X is either strictly increasing or strictly decreasing for all x_i, $i = 1, \ldots, n$, and satisfies for all $i = 1, \ldots, n$,*

$$\frac{\partial h(x)}{\partial x_i} \neq 0 \quad \forall x \in X, \tag{3.9}$$

i.e., $l = 0$, then h is hidden strictly convex.

Proof. If a second order continuously differentiable function h satisfies (3.9), i.e., $l = 0$, then we know from Theorem 3.7 that $h(x)$ is a hidden strictly convex function on X if $P_1 \neq \emptyset$. Because $1 \in NZ$, we can take p_1, satisfying $\operatorname{sign}(p_1) = \operatorname{sign}\left(\dfrac{\partial h(x)}{\partial x_1} \right)$ and $|p_1|$ large enough such that $\alpha_1 + |p_1| m_1 > 0$, resulting in $\nu_{1,p_1}^h > 0$. Thus, if $l = 0$,

then we can have $\nu_1^h > 0$, i.e., $P_1 \neq \emptyset$. Therefore, if $l = 0$, then h is hidden strictly convex. □

Corollary 3.8 reveals that if a second order continuously differentiable strictly monotone function does not have any of its partial derivatives equal to zero at any point on X, then the monotone function is hidden convex. Essentially, this result was proved earlier in [12], which found that such a monotone function is a d-hidden convex function on X.

Let the upper and lower bounds of $\dfrac{\partial h(x)}{\partial x_i}$, η_i and ζ_i, be defined as follows:

$$\eta_i \geq \max_{x \in X} \frac{\partial h(x)}{\partial x_i},$$
$$\zeta_i \leq \min_{x \in X} \frac{\partial h(x)}{\partial x_i}.$$

Recall l as defined in the beginning of subsection 3.2 and γ_l^h as defined in (3.5). Let

$$A_i = \{p_i \in R \mid \zeta_i \cdot p_i \cdot s(p_i) + \eta_i \cdot p_i \cdot s(-p_i) > -\gamma_l^h\}, \tag{3.10}$$
$$A = \{p = (p_1, \ldots, p_l)^T \mid p_i \in A_i,\ i = 1, \ldots, l\},$$

where

$$s(a) = \begin{cases} 1 & a \geq 0, \\ 0 & a < 0. \end{cases}$$

THEOREM 3.9. *Assume that h is second order continuously differentiable on X. If $A \neq \emptyset$, then h must be a hidden strictly convex function on X.*

Proof. For $d \in S^l$ and $x \in X$, we have

$$\begin{aligned}
\nu_{l,p}^h &= \min_{x \in X, d \in S^l, d^T a_{h,l}(x) = 0} d^T \Big(H_l(x) + C_{p,l}(x) \Big) d \\
&\geq \min_{x \in X, d \in S^l, d^T a_{h,l}(x) = 0} d^T H_l(x) d + \min_{x \in X, d \in S^l, d^T a_{h,l}(x) = 0} \sum_{i=1}^{l} p_i \frac{\partial h(x)}{\partial x_i} d_i^2 \\
&\geq \gamma_l^h + \min_{d \in S^l} \sum_{i=1}^{l} \Big(\zeta_i \cdot p_i \cdot s(p_i) + \eta_i \cdot p_i \cdot s(-p_i) \Big) d_i^2 \\
&\geq \gamma_l^h + \min_{1 \leq i \leq l} \Big(\zeta_i \cdot p_i \cdot s(p_i) + \eta_i \cdot p_i \cdot s(-p_i) \Big).
\end{aligned}$$

If $A \neq \emptyset$, i.e., there exists $(p_1, \ldots, p_l)^T \in R^l$, such that for any $i = 1, \ldots, l$, $\zeta_i \cdot p_i \cdot s(p_i) + \eta_i \cdot p_i \cdot s(-p_i) > -\gamma_l^h$, i.e., $\gamma_l^h + \min_{1 \leq i \leq l} (\zeta_i \cdot p_i \cdot s(p_i) + \eta_i \cdot p_i \cdot s(-p_i)) > 0$. Thus, there exists $(p_1, \ldots, p_l)^T \in R^l$, such that $\nu_{l,p}^h > 0$, which implies that h is a hidden strictly convex function based on Theorem 3.7. Thus, if $A \neq \emptyset$, then h must be a hidden strictly convex function. □

Remark 3.1. (i) Recall that a strictly positive γ^h gives an r-hidden strictly convex function. At the same time, if $\gamma^h > 0$, then any A_i is nonempty, as evidenced by $p_i = 0 \in A_i$, $i = 1, \ldots, n$. Thus, any function with $\gamma^h > 0$ is hidden strictly convex. (ii) Recall that a function is d-hidden strictly convex if $\{p_i \mid \zeta_i \cdot p_i \cdot s(p_i) + \eta_i \cdot p_i \cdot s(-p_i) > -\lambda_l^h\}$ is nonempty for all $i = 1, \ldots, n$ [12]. Because $\gamma_l^h \geq \lambda_l^h$, a nonempty $\{p_i \mid \zeta_i \cdot p_i \cdot s(p_i) + \eta_i \cdot p_i \cdot s(-p_i) > -\lambda_l^h\}$ is a subset of A_i. Thus, a function is hidden strictly convex if $\{p_i \mid \zeta_i \cdot p_i \cdot s(p_i) + \eta_i \cdot p_i \cdot s(-p_i) > -\lambda_l^h\}$ is nonempty for all $i = 1, \ldots, n$. In this regard, Theorem 3.9 includes r-hidden convex functions and d-hidden convex function as its special cases.

3.3. Relationship among hidden convexity, monotonicity, pseudoconvexity, and quasiconvexity. We have so far given some sufficient conditions to identify a hidden-convex function. Recall that all convex functions, r-hidden convex functions, d-hidden convex functions, and strictly monotone functions (with all partial derivatives being nonzero at all points in its domain) are hidden-convex functions, and any r-hidden convex function is a pseudoconvex function. We now explore the relationships among these different generalized convex functions.

We first introduce the following notations.

C: Family of second order continuously differentiable convex functions on X.

P: Family of second order continuously differentiable pseudoconvex functions on X.

Q: Family of second order continuously differentiable quasiconvex functions on X.

DH: Family of second order continuously differentiable d-hidden convex functions on X.

RH: Family of second order continuously differentiable r-hidden convex functions on X.

M: Family of second order continuously differentiable strictly monotone functions with all partial derivatives being nonzero at all points on X.

H: Family of second order continuously differentiable hidden-convex functions on X.

The following conclusions can now be obtained.

(1) $C \subset DH \subset H$;
(2) $C \subset RH \subset H$;
(3) $RH \subset P \subset Q$;
(4) $M \subset DH \subset H$;
(5) $M \cup RH$ is a strict subset of H; and
(6) $P \not\subseteq H,\ H \not\subseteq P$, and $H \not\subseteq Q$.

The first four inclusion relationships are apparent from our earlier discussion and from the literature. In the following, we will give some examples to show that $M \cup RH$ is a strict subset of H and that H has no inclusion relationship with either P or Q.

Example 3.1. Let $h(x) = x_1^3 - 4x_1^2 - x_1x_2 - \frac{1}{2}x_2^2 - x_1x_3 - \frac{1}{2}x_3^2 + 10x_3$ and $X = \{(x_1, x_2, x_3)^T \mid 1 \le x_1 \le 5,\ 1 \le x_i \le 2,\ i = 2, 3\}$. Obviously, we have

$$\frac{\partial h(x)}{\partial x_1} = 3x_1^2 - 8x_1 - x_2 - x_3, \quad \frac{\partial h(x)}{\partial x_2} = -x_1 - x_2,$$
$$\frac{\partial h(x)}{\partial x_3} = -x_1 - x_3 + 10.$$

It is easy to check that $NZ = \{2, 3\}$, where NZ is defined by (3.2). Because h is not monotone with respect to x_1, $h \notin M$. The Hessian of $h(x)$ is

$$H(x) = \begin{pmatrix} 6x_1 - 8 & -1 & -1 \\ -1 & -1 & 0 \\ -1 & 0 & -1 \end{pmatrix},$$

with eigenvalues

$$\lambda_1 = \frac{(6x_1 - 7) + \sqrt{(6x_1 - 7)^2 + 8}}{2} - 1,\ \lambda_2 = \frac{(6x_1 - 7) - \sqrt{(6x_1 - 7)^2 + 8}}{2} - 1$$
$$\lambda_3 = -1.$$

It is clear that $H(x)$ has two negative eigenvalues for any $x \in X$, which implies that h is not a pseudoconvex function, as the Hessian of any pseudoconvex function has at most one negative eigenvalue (see Corollary 3.18 in [1]). Thus, it is impossible for h to be an RH function, i.e., $h \notin RH$.

We can prove, however, that h is a hidden strictly convex function. In fact, because $NZ = \{2, 3\}$, $l = 1$, and $\gamma_l^h = \min_{x \in X, \frac{\partial h(x)}{\partial x_1} = 0}(6x_1 - 8) = \min_{x \in X} 2\Big(4 + \sqrt{16 + 3(x_2 + x_3)}\Big) - 8 = 2\sqrt{22} > 0$. Setting $p_1 = 0$ in (3.10) leads to $A_1 = A \neq \emptyset$. Thus, based on Theorem 3.9, h is a hidden strictly convex function on X, i.e., $h \in H$. The above example gives concrete evidence that $M \cup RH$ is a strict subset of H.

Example 3.2. Let $h(x) = \dfrac{x_2}{x_1}$ and $X = \{(x_1, x_2)^T \mid 1 \leq x_1 \leq 2, -1 \leq x_2 \leq 1\}$. Note that for any $\bar{x} \in X$, $\left(\dfrac{\partial h(\bar{x})}{\partial x}\right)^T (x - \bar{x}) = \left(-\dfrac{\bar{x}_2}{(\bar{x}_1)^2}, \dfrac{1}{\bar{x}_1}\right)(x_1 - \bar{x}_1, x_2 - \bar{x}_2)^T \geq 0$ implies $\dfrac{x_2}{x_1} \geq \dfrac{\bar{x}_2}{\bar{x}_1}$, i.e., $h(x) \geq h(\bar{x})$. Thus, based on the definition of pseudoconvexity given in section 3.1 of Chapter 9 of [10], $h(x)$ is a pseudoconvex function on X, i.e., $h(x) \in P$. Because

$$-1 \leq \frac{\partial h(x)}{\partial x_1} = -\frac{x_2}{x_1^2} \leq 1,$$
$$\frac{1}{2} \leq \frac{\partial h(x)}{\partial x_2} = \frac{1}{x_1} \leq 1,$$

the index set defined in (3.2) is $NZ = \{2\}$ and $h \notin M$. As the Hessian $H(x)$ of $h(x)$ is

$$H(x) = \begin{pmatrix} 2\dfrac{x_2}{x_1^3} & -\dfrac{1}{x_1^2} \\ -\dfrac{1}{x_1^2} & 0 \end{pmatrix},$$

h is nonconvex. Note that $l = 1$, $H_1(x) = \dfrac{2x_2}{x_1^3}$, $p_1 \in R$, $C_{p,1}(x) = -\dfrac{p_1 x_2}{x_1^2}$, $a_{h,1}(x) = -\dfrac{x_2}{x_1^2}$,

$$\gamma_1^h = \min_{x \in X, d \in S^1, d^T a_{h,1}(x) = 0} H_1(x) d^2 = 0,$$
$$\nu_1^h = \min_{x \in X, d \in S^1, d^T a_{h,1}(x) = 0} \Big(H_1(x) + C_{p,1}(x)\Big) d^2 = 0.$$

Thus, both P_1 in Theorem 3.7 and A_1 in Theorem 3.9 are empty sets. Neither Theorem 3.7 nor Theorem 3.9 can be used as a sufficient condition to show the hidden convexity of h.

Example 3.3. Let $h(x) = x_1^2 - 3x_1 - x_2^2 - x_3^2 + 5x_2 + x_3 - 3$ and $X = \{(x_1, x_2, x_3)^T \mid 1 \le x_i \le 2,\ i = 1, 2, 3\}$. Obviously, we have

$$\begin{aligned} -1 \le \frac{\partial h(x)}{\partial x_1} &= 2x_1 - 3 \le 1 \quad \text{for any } x \in X, \\ \frac{\partial h(x)}{\partial x_2} &= -2x_2 + 5 \ge 1 \quad \text{for any } x \in X, \\ \frac{\partial h(x)}{\partial x_3} &= -2x_3 + 1 \le -1 \quad \text{for any } x \in X. \end{aligned}$$

Thus, $h \notin M$. Because the Hessian of $h(x)$,

$$H(x) = \begin{pmatrix} 2 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -2 \end{pmatrix}$$

has two negative eigenvalues, h is not a pseudoconvex function. Because $RH \subset P$, $h \notin RH$.

We can prove, however, that h is a hidden strictly convex function. In fact, because $NZ = \{2, 3\}$, $l = 1$ and $\gamma_l^h = \min_{x \in X, \frac{\partial h(x)}{\partial x_1} = 0} 2 = 2 > 0$. Setting $p_1 = 0$ in (3.10) leads to $A_1 = A \ne \emptyset$. Thus, based on Theorem 3.9, h is a hidden strictly convex function on X, i.e., $h \in H$.

Example 3.4. Let $h(x) = -x_1^2 - x_2^2$ and $X = \{(x_1, x_2)^T \mid 1 \le x_i \le 2,\ i = 1, 2\}$. Because the set $\{(x_1, x_2)^T \in X \mid h(x) \le -2\} = \{(x_1, x_2)^T \in X \mid x_1^2 + x_2^2 \ge 2\}$ is not convex, h is not a quasiconvex function on X. Clearly, h is a monotone function on X that satisfies (3.9). Thus, based on Corollary 3.8, h is a hidden-convex function. In summary, $h \in M \subset H$, but $h \notin Q$.

We note here that for quadratic functions, $RH = P$ (see [14]). Thus, for quadratic functions, we have $P \subseteq H$, and we can further conclude from Example 3.3 that P is a strict subset of H. The hidden convexity is a real expansion of pseudoconvexity for quadratic functions. In summary, we have the following relationships for quadratic functions:

$$C \subset RH = P \subset H,\ M \subset DH \subset H, \text{ and } M \cup P \text{ is a strict subset of } H.$$

The above inclusion relationship can be further extended as follows. Recall the definition of parameter l for the index set $NZ = \{l + 1, \ldots, n\}$ in (3.2). Let us introduce the following notations.

C_l: Family of second order continuously differentiable functions with $1 \le l < n$ that satisfy the following condition: For any fixed $(x_{l+1}, \ldots, x_n)$, h is strictly convex on X with respect to $(x_1, \ldots, x_l)$.

RH_l: Family of second order continuously differentiable functions with $1 \le l < n$ that satisfy the following condition: For any fixed $(x_{l+1}, \ldots, x_n)$, h is r-hidden convex on X with respect to $(x_1, \ldots, x_l)$.

P_l: Family of second order continuously differentiable functions with $1 \le l < n$ that satisfy the following condition: For any fixed $(x_{l+1}, \ldots, x_n)$, h is pseudoconvex on X with respect to $(x_1, \ldots, x_l)$.

The following proposition is obvious, and we thus omit its proof.

PROPOSITION 3.10.

1° $\cup_{l\in\{1,\ldots,n-1\}} C_l$ *is a strict subset of* H;
2° $\cup_{l\in\{1,\ldots,n-1\}} RH_l \subset H$; *and*
3° $\cup_{l\in\{1,\ldots,n-1\}} P_l \subset H$ *for quadratic functions.*

4. Hidden-convex programming problem. We consider the mathematical programming problem (P) in this section. The original problem can be converted into the following problem using the transformation in (3.1):

$$\begin{aligned}(E)\qquad &\min\ \psi_0(y) = T_0\left(g_0\left(t(y)\right)\right)\\ &\text{s.t. } \psi_k(y) = T_k\left(g_k\left(t(y)\right)\right) \le T_k(b_k),\ k = 1,\ldots,m,\\ &\qquad y \in Y,\end{aligned}$$

where $Y = \{y \in R^n \mid t(y) \in X\}$.

We have the following theorem for the equivalence between (P) and (E).

THEOREM 4.1. *If* $T_k : k = 0, 1, \ldots, m$, *are strictly increasing,* $t(y)$ *is a one-to-one mapping from* Y *to* X, *and both* t *and* t^{-1} *are continuous, then* y^* *is a global (local) minimizer of problem* (E) *if and only if* $x^* = t(y^*)$ *is a global (local) minimizer of problem* (P).

Proof. The proof can be found in Theorem 3.1 of [16]. □

DEFINITION 4.2. *If there exist strictly increasing functions* $T_0(\cdot)$, $T_1(\cdot)$, $\ldots$, $T_m(\cdot)$ *and domain transformation* $x = t(y)$ *satisfying the conditions in Theorem* 4.1 *such that the programming problem* (E) *is a (strictly) convex programming problem, then the original problem* (P) *is called a hidden (strictly) convex programming problem.*

COROLLARY 4.3. *If problem* (P) *is a hidden-convex programming problem, then any local minimizer of* (P) *must be a global minimizer.*

Proof. By Definition 4.2, we know that problem (E) is a convex programming problem if (P) is a hidden-convex programming problem. By Theorem 4.1, the local optimality of $\bar{x}$ of (P) implies the local optimality of $\bar{y} = t^{-1}(\bar{x})$ in (E). The convexity of (E) further indicates the global optimality of $\bar{y} = t^{-1}(\bar{x})$ in (E), which finally confirms the global optimality of $\bar{x}$ in (P), again by Theorem 4.1. □

Based on Corollary 4.3, if we identify problem (P) to be a hidden-convex programming problem, then we can obtain its global minimizer by solving the original problem (P) using local search methods. There is no need to actually implement a transformation.

In the following, we discuss the conditions for identifying hidden-convex programming problems.

Let η_i^k and ζ_i^k be the upper and lower bounds of $\dfrac{\partial g_k(x)}{\partial x_i}$ on X, respectively, $k = 0, 1, \ldots, m$, $i = 1, \ldots, n$, i.e.,

$$\eta_i^k \ge \max_{x\in X} \frac{\partial g_k(x)}{\partial x_i},$$
$$\zeta_i^k \le \min_{x\in X} \frac{\partial g_k(x)}{\partial x_i}.$$

Let $G_k(z)$ be the Hessian matrix of g_k at z, $k = 0, 1, \ldots, m$. Define

$$\begin{aligned} a_k(x) &= \Big(\frac{\partial g_k(x)}{\partial x_1}, \ldots, \frac{\partial g_k(x)}{\partial x_n}\Big)^T, \\ \lambda^{(k)} &= \min_{x \in X, d \in S^n} d^T G_k(x) d, \\ \gamma^{(k)} &= \min_{x \in X, d \in S^n, a_k^T(x) d = 0} d^T G_k(x) d. \end{aligned}$$

THEOREM 4.4. *Assume that in problem* (P)*, all* g_k*,* $k = 0, 1, \ldots, m$*, are second order continuously differentiable on* X*. For any* $k = 0, 1, \ldots, m,\ i = 1, \ldots, n$*, let*

$$\begin{aligned} A_k^i &= \{p_i \in R \mid \zeta_i^k \cdot p_i \cdot s(p_i) + \eta_i^k \cdot p_i \cdot s(-p_i) > -\gamma^{(k)}\}, \\ A^i &= \cap_{k=0}^m A_k^i, \\ A &= \{p = (p_1, \ldots, p_n)^T \mid p_i \in A^i,\ i = 1, \ldots, n\}. \end{aligned}$$

If $A \neq \emptyset$*, then the original problem* (P) *is a hidden strictly convex programming problem.*

Proof. The condition $A \neq \emptyset$ implies $A_k^i \neq \emptyset$ for any $i = 1, \ldots, n,\ k = 0, 1, \ldots, m$. Thus, for any $i = 1, \ldots, n,\ k = 0, 1, \ldots, m$, there exists a $p_i \in R$ such that $\zeta_i^k \cdot p_i \cdot s(p_i) + \eta_i^k \cdot p_i \cdot s(-p_i) > -\gamma^{(k)}$. If we take a $p = (p_1, \ldots, p_n)^T$ from A and perform the domain transformation $t_{i,p_i}(y_i) = \begin{cases} y_i & p_i = 0 \\ -\frac{1}{p_i} \ln y_i & p_i \neq 0 \end{cases}$ and the range transformations $T_{q,k}(s) = \exp(qs),\ k = 0, 1, \ldots, m,\ q > 0$, then $t_p(y)$ satisfies

$$\begin{aligned} & t'_{i,p_i}(y_i) \neq 0, \\ & \frac{t''_{i,p_i}(y_i)}{\Big[t'_{i,p_i}(y_i)\Big]^2} = p_i \ \text{ for any } y_i \in Y_{i,p},\ i = 1, \ldots, n, \end{aligned}$$

and $T_{q,k}(s),\ k = 0, 1, \ldots, m$, satisfy

$$T'_{q,k}(s) > 0,\ \frac{T''_{q,k}(s)}{T'_{q,k}(s)} = q \quad \forall s \in R.$$

From the proof of Theorem 3.7, we know that when q is large enough, the functions $T_{q,k}(g_k(t_p(y))),\ k = 0, 1, \ldots, m$, are all strictly convex functions on Y. Thus, problem (E) is a strictly convex programming problem on Y. Furthermore, the original problem (P) is a hidden strictly convex programming problem. ∎

Remark 4.1. Because for any $k = 0, 1, \ldots, m,\ \gamma^{(k)} \geq \lambda^{(k)}$, then for any $k = 0, 1, \ldots, m,\ i = 1, \ldots, n$, we have

$$\{p_i \in R \mid \zeta_i^k \cdot p_i \cdot s(p_i) + \eta_i^k \cdot p_i \cdot s(-p_i) > -\lambda^{(k)}\} \subseteq A_k^i.$$

Let

$$\begin{aligned} I_i &= \{k \mid \zeta_i^k > 0,\ k \in \{0, 1, \ldots, m\}\}, \\ J_i &= \{k \mid \eta_i^k < 0,\ k \in \{0, 1, \ldots, m\}\}, \\ \bar{I}_i &= \{0, 1, \ldots, m\} \setminus I_i, \\ \bar{J}_i &= \{0, 1, \ldots, m\} \setminus J_i. \end{aligned}$$

Without loss of generality, we assume that for any $k = 0, 1, \ldots, m,\ i = 1, \ldots, n$, $\zeta_i^k \neq 0$, and $\eta_i^k \neq 0$. Then, for any $i \in \{1, \ldots, n\}$, we have

$$A^i = \left\{ p_i \in R \,\middle|\, \max_{k \in I_i} \left\{ -\frac{\gamma^{(k)}}{\zeta_i^k} \right\} < p_i < \min_{k \in \bar{I}_i} \left\{ -\frac{\gamma^{(k)}}{\zeta_i^k} \right\} \text{ and } p_i \geq 0 \right\}$$
$$\cup \left\{ p_i \in R \,\middle|\, \max_{k \in \bar{J}_i} \left\{ -\frac{\gamma^{(k)}}{\eta_i^k} \right\} < p_i < \min_{k \in J_i} \left\{ -\frac{\gamma^{(k)}}{\eta_i^k} \right\} \text{ and } p_i < 0 \right\}.$$

From Theorem 4.4, we can obtain the following sufficient conditions for hidden-convex programming problems.

THEOREM 4.5. *Assume that for all $k = 0, 1, \ldots, m$, g_k are second order continuously differentiable on X. If for any $i = 1, \ldots, n$, one of the following conditions hold:*

$$\max \left\{ 0, \max_{k \in I_i} \left[-\frac{\gamma^{(k)}}{\zeta_i^k} \right] \right\} < \min_{k \in \bar{I}_i} \left\{ -\frac{\gamma^{(k)}}{\zeta_i^k} \right\} \tag{4.1}$$

$$\text{or} \quad \max_{k \in \bar{J}_i} \left\{ -\frac{\gamma^{(k)}}{\eta_i^k} \right\} < \min \left\{ 0, \min_{k \in J_i} \left[-\frac{\gamma^{(k)}}{\eta_i^k} \right] \right\}, \tag{4.2}$$

then the original problem (P) *is a hidden strictly convex programming problem.*

Proof. If for all $i = 1, \ldots, n$, either (4.1) or (4.2) holds, then for any $i = 1, \ldots, n$, $A^i \neq \emptyset$, which implies $A \neq \emptyset$. Thus, from Theorem 4.4, the original problem (P) is a hidden strictly convex programming problem. □

Because $\gamma^{(k)} \geq \lambda^{(k)}$ for $k = 0, 1, \ldots, m$, condition

$$\max \left\{ 0, \max_{k \in I_i} \left[-\frac{\lambda^{(k)}}{\zeta_i^k} \right] \right\} < \min_{k \in \bar{I}_i} \left\{ -\frac{\lambda^{(k)}}{\zeta_i^k} \right\}$$

leads to the satisfaction of (4.1) and

$$\max_{k \in \bar{J}_i} \left\{ -\frac{\lambda^{(k)}}{\eta_i^k} \right\} < \min \left\{ 0, \min_{k \in J_i} \left[-\frac{\lambda^{(k)}}{\eta_i^k} \right] \right\}$$

leads to the satisfaction of (4.2). Compared with Theorem 3.3 in [12], Theorem 4.5 includes d-hidden convexity as its special case.

Theorem 4.5 offers sufficient conditions to identify whether a programming problem is a hidden strictly convex programming problem or not. These conditions are transformation free, i.e., they can be determined only by the parameters that are derived from the original problem and are independent of the transformations T_i and t. We need to emphasize that, when using Theorem 4.5, the task to estimate the constrained minimum eigenvalue $\gamma^{(k)}$, in general, is very difficult. Without efficient numerical algorithms to determine $\gamma^{(k)}$ for general programming problems, the above results mainly serve as some promising theoretical findings at this stage. In the next section, however, we will show that for quadratic programming problems, the conditions for checking the hidden convexity are readily verifiable.

5. Hidden convex quadratic optimization problems. We consider the following general quadratic optimization problem:

$$(Q) \qquad \min\ g_0(x) = \frac{1}{2} x^T A^{(0)} x + [b^{(0)}]^T x + c^{(0)}$$
$$\text{s.t. } g_k(x) = \frac{1}{2} x^T A^{(k)} x + [b^{(k)}]^T x + c^{(k)} \leq 0,\ k = 1, \ldots, m,$$
$$x \in X = \{(x_1, \ldots, x_n) \mid l_i \leq x_i \leq u_i, i = 1, \ldots, n\},$$

where for $k = 0, 1, \ldots, m$, $A^{(k)} = \{a_{ij}^{(k)}\}_{n\times n}$ is an $n \times n$ symmetric constant matrix and $b^{(k)} = (b_1^{(k)}, \ldots, b_n^{(k)})^T$ is an n-dimensional constant vector. Obviously, matrix $A^{(k)}$ is the Hessian matrix of $g^{(k)}$ and

$$\frac{\partial g_k(x)}{\partial x} = A^{(k)}x + b^{(k)},\ k = 0, 1, \ldots, m.$$

Let $\lambda^{(k)}$ be the minimal eigenvalue of $A^{(k)}$ and $\gamma^{(k)}$ the constrained minimum eigenvalue of $A^{(k)}$, i.e.,

$$\gamma^{(k)} = \min_{x\in X, d\in S^n, (A^{(k)}x+b^{(k)})^T d=0} d^T A^{(k)} d.$$

The constrained minimum eigenvalue $\gamma^{(k)}$ for a constant matrix A can be easily obtained following the procedure suggested in subsection 2.3.

Let

$$\zeta_i^k = \min_{x\in X} \frac{\partial g_k(x)}{\partial x_i} = \min_{x\in X} \left[\sum_{j=1}^n a_{ij}^{(k)} x_j + b_i^{(k)}\right],$$
$$\eta_i^k = \max_{x\in X} \frac{\partial g_k(x)}{\partial x_i} = \max_{x\in X} \left[\sum_{j=1}^n a_{ij}^{(k)} x_j + b_i^{(k)}\right].$$

Note here that the minimum and maximum values of a linear function over a box, ζ_i^k and η_i^k, are easy to calculate.

Define

$$\begin{aligned}
I_i &= \{k \mid \zeta_i^k > 0,\ k = 0, 1, \ldots, m\},\\
\widetilde{I}_i &= \{k \mid \zeta_i^k < 0,\ k = 0, 1, \ldots, m\},\\
I_i^0 &= \{k \mid \zeta_i^k = 0,\ k = 0, 1, \ldots, m\},\\
J_i &= \{k \mid \eta_i^k < 0,\ k = 0, 1, \ldots, m\},\\
\widetilde{J}_i &= \{k \mid \eta_i^k > 0,\ k = 0, 1, \ldots, m\},\\
J_i^0 &= \{k \mid \eta_i^k = 0,\ k = 0, 1, \ldots, m\}.
\end{aligned}$$

From the results in the previous sections, the following lemma and theorem are obvious. We omit their proofs.

LEMMA 5.1. *If $\gamma^{(k)} \geq 0$ for all $k = 0, 1, \ldots, m$, then the quadratic optimization problem (Q) is hidden convex.*

THEOREM 5.2. *If, for any $i \in \{1, \ldots, n\}$, one of the following two conditions hold:*

$$\min_{k\in I_i^0} \gamma^{(k)} \geq 0 \quad \textit{and} \quad \max\left\{0, \max_{k\in I_i}\left[-\frac{\gamma^{(k)}}{\zeta_i^k}\right]\right\} \leq \min_{k\in \widetilde{I}_i}\left\{-\frac{\gamma^{(k)}}{\zeta_i^k}\right\},$$

or

$$\min_{k\in J_i^0} \gamma^{(k)} \geq 0 \quad \textit{and} \quad \max_{k\in \widetilde{J}_i}\left\{-\frac{\gamma^{(k)}}{\eta_i^k}\right\} \leq \min\left\{0, \min_{k\in J_i}\left[-\frac{\gamma^{(k)}}{\eta_i^k}\right]\right\},$$

then the quadratic optimization problem (Q) *is hidden convex.*

The following example illustrates how to check the hidden convexity of problem (Q) step by step.

Example 5.1.

$$\begin{aligned}
&\min g_0(x) = \frac{1}{2}x^T A^{(0)}x + [b^{(0)}]^T x,\\
&\text{s.t. } g_1(x) = \frac{1}{2}x^T A^{(1)}x + [b^{(1)}]^T x - 4200 \le 0,\\
&\qquad x \in X = \left\{ x \in R^3 \mid 4 \le x_1 \le \frac{9}{2},\ 1 \le x_2 \le 2,\ 1 \le x_3 \le 2 \right\},
\end{aligned}$$

where

$$A^{(0)} = \begin{pmatrix} -2 & 0 & 0 \\ 0 & \frac{7}{2} & -\frac{\sqrt{3}}{2} \\ 0 & -\frac{\sqrt{3}}{2} & \frac{5}{2} \end{pmatrix}, A^{(1)} = \begin{pmatrix} -2 & 0 & 0 \\ 0 & \frac{1}{4} & -\frac{3\sqrt{3}}{4} \\ 0 & -\frac{3\sqrt{3}}{4} & -\frac{5}{4} \end{pmatrix},$$

and

$$b^{(0)} = (0, -(4-\sqrt{3}), -(3-\sqrt{3}))^T, \quad b^{(1)} = (1009, -100, 20)^T.$$

Obviously, this problem is not convex.

By the eigenvalue decomposition software in MATLAB, we obtain

$$\Omega_0 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -0.5000 & -0.8660 \\ 0 & -0.8660 & 0.5000 \end{pmatrix},\ \Omega_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0.5000 & -0.8660 \\ 0 & 0.8660 & 0.5000 \end{pmatrix}$$

such that

$$\begin{aligned}
\Omega_0^{-1} A^{(0)} \Omega_0 &= \begin{pmatrix} -2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 4 \end{pmatrix}\\
\Omega_1^{-1} A^{(1)} \Omega_1 &= \begin{pmatrix} -2 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 1 \end{pmatrix}.
\end{aligned}$$

Thus, the eigenvalues of $A^{(0)}$ are $\lambda^{(0)} = \lambda_1^{(0)} = -2$, $\lambda_2^{(0)} = 2$, $\lambda_3^{(0)} = 4$ and the eigenvalues of $A^{(1)}$ are $\lambda^{(1)} = \lambda_1^{(1)} = \lambda_2^{(1)} = -2$, $\lambda_3^{(1)} = 1$.

In the following, we use Theorem 5.2 to prove that the above example problem is hidden convex. First, we use Proposition 2.4 to obtain $\gamma^{(k)}$ for $k = 0, 1$, where

$$\gamma^{(k)} = \min_{x \in X, d \in S^3, [A^{(k)}x + b^{(k)}]^T d = 0} d^T A^{(k)} d.$$

It is clear from (i) of Proposition 2.4 that $\gamma^{(1)} = -2$. The calculation of $\gamma^{(0)}$

involves several steps:

$$\bar{b}^{(0)} = \Omega_0^T \begin{pmatrix} 0 \\ -(4-\sqrt{3}) \\ -(3-\sqrt{3}) \end{pmatrix} = \begin{pmatrix} 0 \\ -2.232 \\ 1.330 \end{pmatrix},$$

$$\varsigma = \left(\frac{\bar{b}_1^{(0)}}{\lambda_1^{(0)}}, \frac{\bar{b}_2^{(0)}}{\lambda_2^{(0)}}, \frac{\bar{b}_3^{(0)}}{\lambda_3^{(0)}} \right)^T = (0, -1.116, 0.3325)^T,$$

$$\bar{X}^{(0)} = \left\{ x \in R^3 \mid \begin{pmatrix} 4 \\ 1 \\ 1 \end{pmatrix} \leq \Omega_0 x - \Omega_0 \varsigma \leq \begin{pmatrix} \frac{9}{2} \\ 2 \\ 2 \end{pmatrix} \right\}.$$

Solving $\{\min x_1 \mid x \in \bar{X}^{(0)}\}$ yields $v(\min x_1 \mid x \in \bar{X}^{(0)}) = 4 > 0$. Thus, there does not exist any $\bar{x} \in \bar{X}^{(0)}$ such that $\bar{x}_1 = 0$. We further construct

$$\bar{Y}^{(0)} = \left\{ y \in R^3 \mid \begin{pmatrix} 4 \\ 1 \\ 1 \end{pmatrix} y_1 \leq \Omega_0 \begin{pmatrix} 1 \\ y_2 \\ y_3 \end{pmatrix} - \Omega_0 \varsigma y_1 \leq \begin{pmatrix} \frac{9}{2} \\ 2 \\ 2 \end{pmatrix} y_1 \right\}.$$

The value of $\max\{|v(\max y_2 \mid y \in \bar{Y}^{(0)})|^2, |v(\min y_2 \mid y \in \bar{Y}^{(0)})|^2\}$ yields $\beta_{12}^{(0)} = 0.9255$ in (v) of Proposition 2.4. From (v) of Proposition 2.4, we have

$$\gamma^{(0)} = \lambda_1^{(0)} + \frac{\lambda_2^{(0)} - \lambda_1^{(0)}}{1 + \beta_{12}^{(0)} \left(\frac{\lambda_2^{(0)}}{\lambda_1^{(0)}} \right)^2} = 0.0774.$$

It is not difficult to verify the following:

$$\zeta_1^0 = -9 \leq \frac{\partial g_0(x)}{\partial x_1} = -2x_1 \leq -8 = \eta_1^0,$$

$$\zeta_2^0 = -\frac{1}{2} \leq \frac{\partial g_0(x)}{\partial x_2} = \frac{7}{2}x_2 - \frac{\sqrt{3}}{2}x_3 - (4-\sqrt{3}) \leq 3 + \frac{\sqrt{3}}{2} = \eta_2^0,$$

$$\zeta_3^0 = -\frac{1}{2} \leq \frac{\partial g_0(x)}{\partial x_3} = \frac{5}{2}x_3 - \frac{\sqrt{3}}{2}x_2 - (3-\sqrt{3}) \leq 2 + \frac{\sqrt{3}}{2} = \eta_3^0,$$

$$\zeta_1^1 = 1000 \leq \frac{\partial g_1(x)}{\partial x_1} = -2x_1 + 1009 \leq 1001 = \eta_1^1,$$

$$\zeta_2^1 = -\frac{399 + 6\sqrt{3}}{4} \leq \frac{\partial g_1(x)}{\partial x_2} = \frac{1}{4}x_2 - \frac{3\sqrt{3}}{4}x_3 - 100 \leq -\frac{398 + 3\sqrt{3}}{4} = \eta_2^1,$$

$$\zeta_3^1 = \frac{35 - 3\sqrt{3}}{2} \leq \frac{\partial g_1(x)}{\partial x_3} = -\frac{5}{4}x_3 - \frac{3\sqrt{3}}{4}x_2 + 20 \leq \frac{75 - 3\sqrt{3}}{4} = \eta_3^1.$$

It is clear that both functions $g_0(x)$ and $g_1(x)$ are not convex, $g_1(x)$ is not pseudoconvex, and $g_0(x)$ is not monotone with respect to x_2 and x_3. Obviously, $I_1 = \{1\}$, $\widetilde{I}_1 = \{0\}$ and $J_2 = \{1\}$, $\widetilde{J}_2 = \{0\}$, $I_3 = \{1\}$, $\widetilde{I}_3 = \{0\}$. The conditions in Theorem 5.2 can be verified to satisfy for all $i = 1, 2, 3$.

When $i = 1$,

$$\begin{aligned}
&\max\left\{0, \max_{k\in I_1}\left[-\frac{\gamma^{(k)}}{\zeta_1^k}\right]\right\}\\
&= \max\left\{0, -\frac{\gamma^{(1)}}{\zeta_1^1}\right\} = \max\left\{0, -\frac{-2}{1000}\right\} = 0.0020\\
&< 0.0086 = -\frac{0.0774}{-9} = -\frac{\gamma_{(0)}}{\zeta_1^0}\\
&= \min_{k\in\tilde{I}_1}\left\{-\frac{\gamma^{(k)}}{\zeta_1^k}\right\}.
\end{aligned}$$

When $i = 2$,

$$\begin{aligned}
&\max_{k\in\tilde{J}_2}\left\{-\frac{\gamma_{(k)}}{\eta_2^k}\right\}\\
&= -\frac{\gamma^{(0)}}{\eta_2^0} = -\frac{0.0774}{3+\frac{\sqrt{3}}{2}} = -0.0200\\
&<= -0.0198 = -\frac{8}{398+3\sqrt{3}} = \min\left\{0, -\frac{-2}{-\frac{398+3\sqrt{3}}{4}}\right\} = \min\left\{0, -\frac{\gamma^{(1)}}{\eta_2^1}\right\}\\
&= \min\left\{0, \min_{k\in J_2}\left[-\frac{\gamma^{(k)}}{\eta_2^k}\right]\right\}.
\end{aligned}$$

When $i = 3$,

$$\begin{aligned}
&\max\left\{0, \max_{k\in I_3}\left[-\frac{\gamma^{(k)}}{\zeta_3^k}\right]\right\}\\
&= \max\left\{0, -\frac{\gamma^{(1)}}{\zeta_3^1}\right\} = \max\left\{0, -\frac{-2}{\frac{35-3\sqrt{3}}{2}}\right\} = 0.1342\\
&< 0.1548 = -\frac{0.0774}{-\frac{1}{2}} = -\frac{\gamma^{(0)}}{\zeta_3^0}\\
&= \min_{k\in\tilde{I}_3}\left\{-\frac{\gamma^{(k)}}{\zeta_3^k}\right\}.
\end{aligned}$$

We can thus conclude by Theorem 5.2 that Example 5.1 is hidden convex. The local minimizer of this problem, $x^* = (4.2760, 1.1421, 1.0000)^T$ with objective value $f^* = -19.5984$, which is obtained by the constrained nonlinear minimization function *fmincon* of the MATLAB optimization toolbox, must also be the global minimizer.

We present in the following an algorithm for checking the hidden convexity of problem (Q).

ALGORITHM II (Verification of the hidden convexity of problem (Q)).

Step 1. Calculate $\gamma^{(k)} = \min_{x\in X, d\in S^n, (A^{(k)}x+b^{(k)})^T d=0} d^T A^{(k)} d$, $k = 0, 1, \ldots, m$, by Algorithm I. Let $i = 1$.

Step 2. If $i > n$, then go to *Step* 5. Otherwise, for $k = 0, 1, \ldots, m$, calculate ζ_i^k and η_i^k. Let $\Gamma_i := \min_{k\in I_i^0} \gamma^{(k)}$ and $\Upsilon_i := \min_{k\in J_i^0} \gamma^{(k)}$ and go to *Step* 3.

Step 3. Let $\Lambda_i := \max\{0, \max_{k\in I_i}[-\frac{\gamma^{(k)}}{\zeta_i^k}]\}$, $\bar{\Lambda}_i := \min_{k\in\tilde{I}_i}\{-\frac{\gamma^{(k)}}{\zeta_i^k}\}$. If $\Gamma_i \geq 0$ and $\Lambda_i \leq \bar{\Lambda}_i$, then let $i := i + 1$ and go to *Step* 2. Otherwise, go to *Step* 4.

Step 4. Let $\Lambda_i := \max_{k\in\widetilde{J}_i}\{-\frac{\gamma_i^{(k)}}{\eta_i^k}\}$, $\bar{\Lambda}_i := \min\{0, \min_{k\in J_i}[-\frac{\gamma_i^{(k)}}{\eta_i^k}]\}$. If $\Upsilon_i \geq 0$ and $\Lambda_i \leq \bar{\Lambda}_i$, then let $i := i+1$ and go to *Step* 2. Otherwise, stop, and the hidden convexity of problem (Q) cannot be proved.

Step 5. Problem (Q) is hidden convex. Use an appropriate routine to find a local optimal solution of problem (Q), which is also a global optimal solution.

Remark 5.1. Using Algorithms I and II, we can verify whether problem (Q) is a hidden-convex programming problem by solving some linear programming problems. Note that the calculation of ζ_i^k and η_i^k can be simply carried out by determining whether x_j, $j = 1, \ldots, n$, takes its upper bound or its lower bound based on the positiveness or negativeness of the corresponding coefficient $a_{ij}^{(k)}$, $i = 1, \ldots, n$, $k = 0, \ldots, m$. The largest number of actual linear programming problems that we need to solve in Algorithms I and II is $4(m+1)$ and each linear programming problem involves n variables and $2n$ inequality constraints. The complexity analysis of linear programming has been well established and can be applied to estimate the worst-case complexity of the verification process for hidden convexity, given the size of the problem under investigation.

Algorithms I and II are coded in MATLAB 7.0 with the Windows XP operating system, and the Optimization Toolbox 3.0 is used in the coding. We demonstrate in the following how to use Algorithms I and II to check the hidden convexity for a 10-dimensional quadratic optimization problem.

Example 5.2.

$$\begin{aligned}
&\min g_0(x) = \frac{1}{2}x^T A^{(0)} x + [b^{(0)}]^T x\\
&\text{s.t. } g_1(x) = \frac{1}{2}x^T A^{(1)} x + [b^{(1)}]^T x - 1800 \leq 0,\\
&\qquad x \in X = \{x \in R^3 \mid 4 \leq x_1 \leq \frac{9}{2},\ 1 \leq x_i \leq 2,\ i = 2, \ldots, 10\},
\end{aligned}$$

where

$$A^{(0)} = \begin{pmatrix}
-2 & 0 & 0 & -\frac{1}{2} & 0 & 0 & 0 & 0 & 0 & -\frac{1}{2}\\
0 & \frac{7}{2} & -\frac{\sqrt{3}}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & -\frac{\sqrt{3}}{2} & \frac{5}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
-\frac{1}{2} & 0 & 0 & 2 & -\frac{1}{2} & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & -\frac{1}{2} & 2 & -\frac{1}{2} & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & -\frac{1}{2} & 2 & -\frac{1}{2} & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & -\frac{1}{2} & 2 & -\frac{1}{2} & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{2} & 2 & -\frac{1}{2} & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{2} & 2 & -\frac{1}{2}\\
-\frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{2} & 2
\end{pmatrix},$$

$$A^{(1)} = \begin{pmatrix}
-2 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1\\
0 & \frac{1}{4} & -\frac{3\sqrt{3}}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & -\frac{3\sqrt{3}}{4} & -\frac{5}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 1 & -1 & 1 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 1 & -1 & 1 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 1 & -1 & 1 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 1 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 1\\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1
\end{pmatrix},$$

and

$$b^{(0)} = (2, -(4-\sqrt{3}), -(3-\sqrt{3}), -1, -3, -3, -3, -3, -3, -1)^T,$$

$$b^{(1)} = (300, -50, 50, 50, 40, 60, 40, 50, 60, 50)^T.$$

Obviously, this problem is not convex. We first obtain $\gamma^{(0)} = 0.2328$ and $\gamma^{(1)} = -2.8659$ by using Algorithm I. Implementing Algorithm II, we find out that for all $i = 1, \ldots, 10$, $\Gamma_i = \Upsilon_i = \infty$, $[\Lambda_1, \ldots, \Lambda_{10}]^T = [0.0098, -0.0602, 0.0638, 0.0541, 0.0716, 0.0478, 0.0716, 0.0573, 0.0478, 0.0541]^T$ and $[\bar{\Lambda}_1, \ldots, \bar{\Lambda}_{10}]^T = [0.0259, -0.0564, 0.4657, 0.1035, 0.0776, 0.0776, 0.0776, 0.0776, 0.0776, 0.1035]^T$. Thus, we can conclude that Example 5.2 is hidden convex. The global minimizer of this problem is $x^* = (4.3827, 1.2917, 1.0000, 1.2577, 1.6279, 1.4911, 1.7009, 1.5632, 1.3535, 1.1918)^T$ with objective value $f^* = -34.2711$. The calculation is carried out on an Intel Pentium M with 1.60GHz and 512MB RAM and the total computational time is 2.053 seconds.

Remark 5.2. Let (Q_S) be a class of quadratic optimization problems with a single constraint ($m = 1$) and $X = R^n$. The solution properties of (Q_S) were discussed in [6] and [9] under an assumption that Slater's constraint qualification is satisfied. Appendix B of [6] pointed out that the strong duality holds for (Q_S) and its Lagrangian dual problem is a convex programming problem. Using the same notations as in (Q), [9] pointed out that $\bar{x}$ is a global minimizer of problem (Q_S) if and only if there exists a $\lambda \geq 0$ such that

$$A^{(0)} + \lambda A^{(1)} \succeq 0, b^{(0)} + \lambda b^{(1)} + (A^{(0)} + \lambda A^{(1)})\bar{x} = 0, \lambda g_1(\bar{x}) = 0.$$

Although many instances of (Q_S) are not hidden convex, a global minimizer of any problem (Q_S) that satisfies Slater's condition can always be characterized by a convex programming problem. A key point to be emphasized is that any nonconvex optimization problem with multiple minima cannot be hidden convex. Note that if, in (Q_S), X is bounded, as we are discussing in this paper, then the results that are derived in [6] and [9] will not hold true.

6. Conclusion. Transformation-independent sufficient conditions have been developed in this paper to peel off a nonconvex cover of certain actual convex programming problems. The reach of convex analysis can be thus extended to a class of hidden-convex functions. Most promisingly, some implementable checking procedures have been derived in this paper to identify hidden convex quadratic optimization problems.

One future research subject is to integrate the checking procedure for hidden convexity into a branch-and-bound framework to solve a class of global optimization problems. Let us consider the following simple example.

Example 6.1.

$$\begin{aligned} \text{(PE)} \qquad & \min g_0(x) = \frac{1}{2}x^2 + \frac{1}{2}x \\ & \text{s.t. } g_1(x) = -\frac{1}{2}x^2 - 2x - \frac{1}{2} \leq 0, \\ & x \in X = \{x \mid -1 \leq x \leq 1\}. \end{aligned}$$

It is easy to check for this problem that $\gamma^{(0)} = 1$, $\gamma^{(1)} = -1$, $\zeta^0 = \min_{x \in X}(x + \frac{1}{2}) = -\frac{1}{2}$, $\eta^0 = \max_{x \in X}(x + \frac{1}{2}) = \frac{3}{2}$, $\zeta^1 = \min_{x \in X}(-x - 2) = -3$, and $\eta^1 = \max_{x \in X}(-x - 2) = -1$. We further have $I = \emptyset$, $\tilde{I} = \{0, 1\}$, $J = \{1\}$, and $\tilde{J} = \{0\}$. Neither of the

two conditions in Theorem 5.2 is satisfied and we are not able to verify the hidden convexity of problem (PE).

We consider two subregions of X, $X_1 = \{x \mid -1 \le x \le 0\}$ and $X_2 = \{x \mid -\frac{1}{4} \le x \le 1\}$, such that $X = X_1 \cup X_2$. Consider two revised problems of (PE), (PE_1), and (PE_2), by replacing X with X_1 and X_2, respectively. For (PE_1), $\zeta^0 = -\frac{1}{2}$, $\eta^0 = \frac{1}{2}$, $\zeta^1 = -2$, $\eta^1 = -1$, $I = \emptyset$, $\tilde{I} = \{0, 1\}$, $J = \{1\}$, and $\tilde{J} = \{0\}$. For (PE_2), $\zeta^0 = \frac{1}{4}$, $\eta^0 = \frac{3}{2}$, $\zeta^1 = -3$, $\eta^1 = -\frac{7}{4}$, $I = \{0\}$, $\tilde{I} = \{1\}$, $J = \{1\}$, and $\tilde{J} = \{0\}$. For both problems (PE_1) and (PE_2), the second condition in Theorem 5.2 is satisfied and both problems (PE_1) and (PE_2) are thus hidden convex. Comparing the minimizer of problem (PE_1), $x_1^* = -0.2680$ with $g_0(x_1^*) = -0.0980$, and the minimizer of problem (PE_2), $x_2^* = -\frac{1}{4}$ with $g_0(x_2^*) = -0.0938$, we can conclude that $x_1^* = -0.2680$ is the global minimizer of problem (PE).

To extend the implementable computational procedure for checking hidden convexity in nonquadratic optimization situations, we will explore some possibilities in the near future for applying the results in the literature on eigenvalue bounds for interval matrices (see [7], [13]) to calculate a lower bound of the minimum eigenvalue or the constrained minimum eigenvalue for nonconstant matrices.

Acknowledgments. The authors are grateful to two anonymous referees for their extremely valuable comments and suggestions, which have contributed to the significant improvement of this paper.

REFERENCES

[1] M. Avriel, *Nonlinear Programming, Analysis, and Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1976.

[2] M. Avriel, W. E. Diewert, S. Schaible, and I. Zang, *Generalized Concavity*, Plenum Press, New York, 1988.

[3] M. Avriel and S. Schaible, *Second order characterizations of pseudoconvex functions*, Math. Programming, 14 (1978), pp. 170–185.

[4] M. Avriel and I. Zang, *Generalized convex functions with applications to nonlinear programming*, in Mathematical Programs for Activity Analysis, P. van Moeseke, ed., North–Holland, Amsterdam, 1974, pp. 23–33.

[5] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms*, 2nd ed., John Wiley & Sons, New York, 1993.

[6] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2005.

[7] D. Hertz, *The extreme eigenvalues and stability of real symmetric interval matrices,* IEEE Trans. Automat. Control, 37 (1992), pp. 532–535.

[8] R. Horst, *On the convexification of nonlinear programming problems: An applications-oriented survey,* European J. Oper. Res., 15 (1984), pp. 382–392.

[9] V. Jeyakumar, A. M. Rubinov, and Z. Y. Wu, *Nonconvex Quadratic Minimization Problems with Quadratic Constraints: Global Optimality Conditions*, preprint.

[10] O. L. Mangasarian, *Nonlinear Programming*, McGraw-Hill, New York, 1969.

[11] D. Li, X. L. Sun, M. P. Biswal, and F. Gao, *Convexification, concavification, and monotonization in global optimization,* Ann. Oper. Res., 105 (2001), pp. 213–226.

[12] D. Li, Z. Y. Wu, H. W. J. Lee, X. M. Yang, and L. S. Zhang, *Hidden convex minimization*, J. Global Optim., 31 (2005), pp. 211–233.

[13] J. Rohn, *Bounds on eigenvalues of interval matrices,* ZAMM, Z. Angew. Math. Mech., 78 (1998), S1049–S1050.

[14] S. Schaible and M. Avriel, *Second-order characterizations of pseudoconvex quadratic functions*, J. Optim. Theory Appl., 21 (1977), pp. 15–26.

[15] S. Schaible and I. Zang, *On the convexifiability of pseudoconvex C^2-Functions*, Math. Program., 19 (1980), pp. 289–299.

[16] X. L. Sun, K. McKinnon, and D. Li, *A convexification method for a class of global optimization problems with applications to reliability optimization*, J. Global Optim., 21 (2001), pp. 185–199.

SIAM J. OPTIM.
Vol. 18, No. 2, pp. 537–555

USING SAMPLING AND SIMPLEX DERIVATIVES IN PATTERN SEARCH METHODS*

A. L. CUSTÓDIO† AND L. N. VICENTE‡

Abstract. In this paper, we introduce ways of making a pattern search more efficient by reusing previous evaluations of the objective function, based on the computation of simplex derivatives (e.g., simplex gradients). At each iteration, one can attempt to compute an accurate simplex gradient by identifying a sampling set of previously evaluated points with good geometrical properties. This can be done using only past successful iterates or by considering all past function evaluations. The simplex gradient can then be used to reorder the evaluations of the objective function associated with the directions used in the poll step or to update the mesh size parameter according to a sufficient decrease criterion, neither of which requires new function evaluations. We present these procedures in detail and apply them to a set of problems from the CUTEr collection. Numerical results show that these procedures can enhance significantly the practical performance of pattern search methods.

Key words. derivative-free optimization, pattern search methods, simplex gradient, poll ordering, multivariate polynomial interpolation, poisedness

AMS subject classifications. 65D05, 90C30, 90C56

DOI. 10.1137/050646706

1. Introduction.

We are interested in this paper in designing efficient (derivative-free) pattern search methods for nonlinear optimization problems. We focus our attention on unconstrained optimization problems of the form $\min_{x\in\mathbb{R}^n} f(x)$.

The curve representing the objective function value as a function of the number of function evaluations frequently exhibits an L-shape for pattern search runs. This class of methods, perhaps because of their directional features, is relatively good at quickly decreasing the objective function from its initial value. However, they can be slow thereafter and especially towards stationarity, when the frequency of unsuccessful iterations tends to increase.

There has not been much effort in trying to develop efficient serial implementations of pattern search methods for the minimization of general functions. Some attention has been paid to parallelization (see Hough, Kolda, and Torczon [14]). In the context of generating set search methods, Frimannslund and Steihaug [12] rotated the generating sets based on curvature information extracted from function values. Other authors have considered particular instances where the problem structure can be exploited efficiently. Price and Toint [20] examined how to take advantage of partial separability. Alberto *et al.* [2] have shown ways of incorporating user-provided function evaluations. Abramson, Audet, and Dennis [1] looked at the case where some incomplete form of gradient information is available.

*Received by the editors December 4, 2005; accepted for publication (in revised form) January 17, 2007; published electronically May 29, 2007.
http://www.siam.org/journals/siopt/18-2/64670.html

†Departamento de Matemática, FCT-UNL, Quinta da Torre 2829-516 Caparica, Portugal (alcustodio@fct.unl.pt). Support for this author was provided by Centro de Matemática da Universidade de Coimbra, Centro de Matemática e Aplicações da Universidade Nova de Lisboa, Fundação Calouste Gulbenkian, and by FCT under grant POCI/MAT/59442/2004.

‡Departamento de Matemática, Universidade de Coimbra, 3001-454 Coimbra, Portugal (lnv@mat.uc.pt). Support for this author was provided by Centro de Matemática da Universidade de Coimbra and by FCT under grant POCI/MAT/59442/2004.

The goal of this paper is to develop strategies for improving the efficiency of the current pattern search iteration, based on function evaluations obtained at previous iterations. We make no use of or assumption about the structure of the objective function, so that one can apply the techniques here to any functions (in particular, those resulting from running black-box codes or performing physical experiments). More importantly, these strategies (i) require no extra function evaluation and (ii) do not interfere with existing requirements for global convergence.

The paper is organized as follows. Section 2 describes the pattern search framework over which we introduce the material of this paper. Section 3 summarizes geometrical features of sample sets (Λ–poisedness) and simplex derivatives, such as simplex gradients and simplex Hessians.

The key ideas of this paper are reported in section 4, where we show how to use sample sets of points previously evaluated in a pattern search to compute simplex derivatives. The sample sets can be built by storing points where the function has been evaluated or by storing only points which lead to a decrease. The main destination of this computation is the efficient ordering of the directions used for polling. In fact, a descent indicator direction (like a negative simplex gradient) can be used to order the polling directions according to a simple angle criterion.

In section 5 we describe one way of ensuring sample sets with adequate geometry at iterations succeeding unsuccessful ones. We study the pruning properties of negative simplex gradients in section 6. Other uses of simplex derivatives in a pattern search are suggested in section 7, namely, one way of updating the mesh size parameter according to a sufficient decrease condition.

These ideas were tested in a set of CUTEr [13] unconstrained problems, collected from papers on derivative-free optimization. The corresponding numerical results are reported in section 8 and show the effectiveness of using sampling-based simplex derivatives in pattern search. Section 9 states some concluding remarks and ideas for future work. The default norms used in this paper are Euclidean.

2. Pattern search. Pattern search methods are directional methods that make use of a finite number of directions with appropriate descent properties. In the unconstrained case, these directions must positively span $\mathbb{R}^n$. A positive spanning set is guaranteed to contain one positive basis, but it can contain more. A positive basis is a positive spanning set which has no proper subset positively spanning $\mathbb{R}^n$. Positive bases have between $n+1$ and $2n$ elements. Properties and examples of positive bases can be found in [2, 10, 17]. If the objective function possesses certain smoothness properties and the number of positive bases used remains finite, then the pattern search is known to exhibit global convergence to stationary points in the lim inf sense (see [3, 17]).

We present pattern search methods in the generalized format introduced by Audet and Dennis [3]. The positive spanning set used is represented by D and its cardinality by $|D|$. It is convenient to regard D as an $n \times |D|$ matrix whose columns are the elements of D. A positive basis in D is denoted by B and is also viewed as a matrix (an $n \times |B|$ column submatrix of D).

At each iteration k of a pattern search method, the next iterate x_{k+1} is selected among the points of a mesh M_k, defined as

$$M_k = \{x_k + \alpha_k D z : \ z \in \mathbb{Z}_+^{|D|}\},$$

where $\mathbb{Z}_+$ is the set of nonnegative integers. This mesh is centered at the current iterate x_k, and its fineness is defined by the mesh size (or step size) parameter $\alpha_k > 0$.

Each direction $d \in D$ must be of the form $d = G\bar{z},\ \bar{z} \in \mathbb{Z}^n$, where G is a nonsingular (generating) matrix. This property is crucial for global convergence, ensuring that the mesh has only a finite number of points in a compact set (provided that the mesh size parameter is also updated according to some rationality requirements, as we will point out later).

The process of finding a new iterate $x_{k+1} \in M_k$ can be described in two phases (the search step and the poll step). The search step is optional and unnecessary for the convergence properties of the method. It consists of evaluating the objective function at a finite number of points lying on the mesh M_k. The choice of points in M_k is totally flexible as long as its number remains finite. The points could be chosen according to specific application properties or following some heuristic algorithm. The search step is declared successful if a new mesh point x_{k+1} is found such that $f(x_{k+1}) < f(x_k)$.

The poll step is performed only if the search step has been unsuccessful. It consists of a local search around the current iterate, exploring the points in the mesh neighborhood defined by the parameter α_k and a positive basis $B_k \subset D$:

$$P_k \ = \ \{x_k + \alpha_k b : \ b \in B_k\} \ \subset \ M_k.$$

We call the points $x_k + \alpha_k b \in P_k$ the polling points and the vectors $b \in B_k$ the polling vectors or polling directions.

The purpose of the poll step is to ensure a decrease in the objective function for sufficiently small mesh sizes. Provided that the function retains some differentiability properties, one knows that the poll step must be eventually successful, unless the current iterate is a stationary point. In fact, given any vector w in $\mathbb{R}^n$, there exists at least one vector b in B_k such that $w^\top b \ > 0$ (see [10]). For instance, if the function f is continuously differentiable and one selects $w = -\nabla f(x_k)$, then one is guaranteed the existence of a descent direction in B_k.

The polling vectors (or points) are ordered according to some criterion in the poll step. The report [18] presents two distinct classes of pattern search algorithms, namely, the rank ordering and the positive bases pattern search methods. In the context of a rank ordering pattern search, it is suggested to order the simplex vertices but with the single purpose of identifying the vertices with the best and the worst objective function values in order to compute a crude estimate of the direction of steepest descent. The authors explicitly state in [18] that their intention was not reordering the remaining vertices. Most papers do not address the issue of poll ordering at all and, as a result, numerical testing is typically done using the ordering in which the vectors are originally stored. Another ordering we discuss later consists of bringing into the first column (in B_{k+1}) the polling vector b_k associated with the most recent successful polling iterate ($f(x_k + \alpha_k b_k) < f(x_k)$). This ordering procedure has been called *dynamic polling* (see Audet and Dennis [4]). Our presentation of a pattern search assumes that poll ordering is specified before polling starts.

If the poll step also fails to produce a point with a lower objective function value $f(x_k)$, then both the poll step and the iteration are declared unsuccessful. In this situation the mesh size parameter is decreased. On the other hand, the mesh size is held constant or increased if, in either the search or the poll step, a new iterate is found yielding an objective function decrease.

The class of pattern search methods used in this paper is described in Figure 2.1. Our description follows the one given in [3] for the generalized pattern search. We leave three procedures undetermined in the statement of the method: the search procedure in the search step, the order procedure that determines the order of the

Pattern Search Method

Initialization
Choose x_0 and $\alpha_0 > 0$. Choose a positive spanning set D. Select all constants needed for procedures `[search]`, `[order]`, and `[mesh]`. Set $k = 0$.

Search step
Call `[search]` to try to compute a point $x \in M_k$ with $f(x) < f(x_k)$ by evaluating the function only at a finite number of points in M_k. If such a point is found, then set $x_{k+1} = x$, declare the iteration as successful, and skip the poll step.

Poll step
Choose a positive basis $B_k \subset D$. Call `[order]` to order the polling set $P_k = \{x_k + \alpha_k b : b \in B_k\}$. Start evaluating f at the polling points following the order determined. If a polling point $x_k + \alpha_k b_k$ is found such that $f(x_k + \alpha_k b_k) < f(x_k)$, then stop polling, set $x_{k+1} = x_k + \alpha_k b_k$, and declare the iteration as successful. Otherwise declare the iteration as unsuccessful and set $x_{k+1} = x_k$.

Updating the mesh size parameter
Call `[mesh]` to compute α_{k+1}. Increment k by one and return to the search step.

FIG. 2.1. *Class of pattern search methods used in this paper.*

`procedure mesh`

The constant τ must satisfy $\tau \in \mathbb{Q}$ and $\tau > 1$ and should be initialized at iteration $k = 0$ together with $j_{max} \in \mathbb{Z}$, $j_{max} \geq 0$, and $j_{min} \in \mathbb{Z}$, $j_{min} \leq -1$.

If the iteration was successful, then maintain or expand the mesh by taking $\alpha_{k+1} = \tau^{j_k^+} \alpha_k$, with $j_k^+ \in \{0, 1, 2, \ldots, j_{max}\}$. Otherwise, contract the mesh by decreasing the mesh size parameter $\alpha_{k+1} = \tau^{j_k^-} \alpha_k$, with $j_k^- \in \{j_{min}, \ldots, -1\}$.

FIG. 2.2. *Updating the mesh size parameter (for rational lattice requirements).*

polling directions, and the mesh procedure that updates the mesh size parameter. These procedures are called within squared brackets for better visibility.

The search and order routines are not asked to meet any requirements for global convergence purposes (except for finiteness of the number of mesh points considered in search).

The mesh procedure, however, must update the mesh size parameter as described in Figure 2.2. The most common choice is to divide the parameter in half at unsuccessful iterations and to keep it or double it at successful ones. As noted by Hough, Kolda, and Torczon [14], increasing the mesh size parameter for all successful iterations can result in an excessive number of later contractions, each one requiring a complete polling, thus leading to an increase in the total number of function evaluations required. A possible strategy to avoid this behavior (fitting the procedure of Figure 2.2) has been suggested in [14] and consists of expanding the mesh only if two consecutive successful iterates have been computed using the same direction.

The global convergence analysis for this class of pattern search methods is divided into two parts. The first part establishes that a subsequence of mesh size parameters goes to zero. This result was first proved by Torczon in [21], and it is stated here as Theorem 2.1.

THEOREM 2.1. *Consider a sequence* $\{x_k\}$ *of pattern search iterates. If* $L(x_0) = \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ *is compact, then the sequence of the mesh size parameters satisfies* $\liminf_{k \to +\infty} \alpha_k = 0$.

The second part of the analysis requires some differentiability properties of the objective function and can be found, for instance, in [3, 17]. We formalize it here for unconstrained minimization.

THEOREM 2.2. *Consider a sequence* $\{x_k\}$ *of pattern search iterates. If* $L(x_0) = \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ *is compact, then there exists at least one convergent subsequence* $\{x_k\}_{k \in K}$ *(with limit point* x_**) of unsuccessful iterates for which the corresponding subsequence of the mesh size parameters* $\{\alpha_k\}_{k \in K}$ *converges to zero. If* f *is strictly differentiable near* x_**, then* $\nabla f(x_*) = 0$. *If* f *is continuously differentiable in an open set containing* $L(x_0)$*, then* $\liminf_{k \longrightarrow +\infty} \|\nabla f(x_k)\| = 0$.

Pattern search and direct search methods for unconstrained optimization are surveyed in the comprehensive paper of Kolda, Lewis, and Torczon [17].

3. Simplex derivatives. A simplex derivative of order one is known as a *simplex gradient*. Simplex gradients were used by Bortz and Kelley [5] in their implicit filtering method, which can be viewed as a line search method based on simplex gradients. Tseng [22] developed a class of simplex-based direct search methods imposing sufficient decrease conditions. He suggested the use of the norm of a simplex gradient in a stopping criterion for his class of methods. No numerical results were reported with this criterion, and no other use of the simplex gradient was suggested. In the context of the Nelder–Mead simplex-based direct search algorithm, Kelley [15] used the simplex gradient norm in a sufficient decrease-type condition to detect stagnation, and the simplex gradient signs to orient the simplex restarts.

Calculation of a simplex gradient first requires the selection of a set of sample points. The geometrical properties of the sample set determine the quality of the corresponding simplex gradient as an approximation to the exact gradient of the objective function. In this paper, we use (determined) simplex gradients as well as underdetermined and overdetermined (or regression) simplex gradients.

In the determined case, a simplex gradient is computed by first sampling the objective function at $n+1$ points. The convex hull of a set of $n+1$ affinely independent points $\{y^0, y^1, \ldots, y^n\}$ is called a simplex. The $n+1$ points are called the vertices of the simplex. Since the points are affinely independent, the matrix $S = [\, y^1 - y^0 \ \ldots \ y^n - y^0 \,]$ is nonsingular. Given a simplex of vertices $y^0, y^1, \ldots, y^n$, the simplex gradient at y^0 is defined as $\nabla_s f(y^0) = S^{-\top} \delta(f; S)$, with $\delta(f; S) = [\, f(y^1) - f(y^0), \ldots, f(y^n) - f(y^0) \,]^\top$.

The simplex gradient is intimately related to linear multivariate polynomial interpolation. In fact, it is easy to see that the linear model $m(y) = f(y^0) + \nabla_s f(y^0)^\top (y - y^0)$ centered at y^0 interpolates f at the points $y^1, \ldots, y^n$.

In practical instances, one might have $q+1 \neq n+1$ points from which to compute a simplex gradient. We say that a sample set is *poised* for a simplex gradient calculation if S is full rank, i.e., if $\operatorname{rank}(S) = \min\{n, q\}$. (The notions of poisedness and affine independence coincide for $q \leq n$, but affine independence is not defined when $q > n$.) Given the sample set $\{y^0, y^1, \ldots, y^q\}$, the simplex gradient $\nabla_s f(y^0)$ of f at y^0 can be defined as the "solution" g of the system

$$S^\top g \ = \ \delta(f; S),$$

where $S = [\, y^1 - y^0 \ \ldots \ y^q - y^0 \,]$ and $\delta(f; S) = [\, f(y^1) - f(y^0), \ldots, f(y^q) - f(y^0) \,]^\top$. This system is solved in the least-squares sense if $q > n$. A minimum norm solution is computed if $q < n$. This definition includes the determined case ($q = n$) as a particular case.

The formulas for the nondetermined simplex gradients can be expressed using the reduced singular value decomposition (SVD) of $S^\top$. However, to deal with the geometrical properties of the poised sample set and to better express the error bound for the corresponding gradient approximation, it is appropriate to take the reduced SVD of a scaled form of $S^\top$. For this purpose, let

$$\Delta \ = \ \max_{1 \leq i \leq q} \|y^i - y^0\|,$$

which is the radius of the smallest enclosing ball of $\{y^0, y^1, \ldots, y^q\}$ centered at y^0. Now we write the reduced SVD of the scaled matrix $S^\top/\Delta = U\Sigma V^\top$, which corresponds to a sample set in a ball of radius one centered around y^0. The underdetermined and overdetermined simplex gradients are both given by $\nabla_s f(y^0) = V\Sigma^{-1}U^\top \delta(f;S)/\Delta$.

The accuracy of simplex gradients is summarized in the following theorem. The proof of the determined case ($q = n$) is given, for instance, by Kelley [16]. The extension of the analysis to the nondetermined cases is developed by Conn, Scheinberg, and Vicente [6].

THEOREM 3.1. *Let $\{y^0, y^1, \ldots, y^q\}$ be a poised sample set for a simplex gradient calculation in $\mathbb{R}^n$. Consider the enclosing (closed) ball $\mathcal{B}(y^0;\Delta)$ of this sample set, centered at y^0, where $\Delta = \max_{1 \leq i \leq q} \|y^i - y^0\|$. Let $S = [\, y^1 - y^0 \ \ldots \ y^q - y^0 \,]$, and let $U\Sigma V^\top$ be the reduced SVD of $S^\top/\Delta$.*

Assume that ∇f is Lipschitz continuous in an open domain Ω containing $\mathcal{B}(y^0;\Delta)$ with constant $\gamma > 0$.

Then the error of the simplex gradient at y^0, as an approximation to $\nabla f(y^0)$, satisfies

$$\|\hat{V}^\top [\nabla f(y^0) - \nabla_s f(y^0)]\| \ \leq \ \left(q^{\frac{1}{2}} \frac{\gamma}{2} \|\Sigma^{-1}\| \right) \Delta,$$

where $\hat{V} = I$ if $q \geq n$ and $\hat{V} = V$ if $q < n$.

Notice that the error difference is projected over the null space of $S^\top/\Delta$. Unless we have enough points ($q + 1 \geq n + 1$), there is no guarantee of accuracy for the simplex gradient. Despite this observation, underdetermined simplex gradients contain relevant gradient information for q close to n and might be of some value in computations where the number of sample points is relatively low.

The quality of the error bound of Theorem 3.1 depends on the size of the constant $\sqrt{q}\gamma\|\Sigma^{-1}\|/2$, which multiplies Δ. This constant, in turn, depends essentially on an unknown Lipschitz constant γ and on $\|\Sigma^{-1}\|$, which is associated to the geometry of the sample set.

Conn, Scheinberg, and Vicente [7] introduced an algorithmic framework for building and maintaining sample sets with good geometry. They have suggested the notion of a Λ–poised sample set, where Λ is a positive constant. The notion of Λ–poisedness is closely related to Lagrange interpolation [7, 6]. If a sample set $\{y^0, y^1, \ldots, y^q\}$ is Λ–poised in the sense of [7, 6], then one can prove that $\|\Sigma^{-1}\|$ is bounded by a multiple of Λ. For the purpose of this paper, it is enough to consider $\|\Sigma^{-1}\|$ as a measure of the well-poisedness (quality of the geometry) of our sample sets. We therefore say that a poised sample set is Λ–*poised* if $\|\Sigma^{-1}\| \leq \Lambda$, for some positive constant Λ.

In a pattern search, we do not necessarily need an algorithm to build or maintain Λ–poised sets. Rather, we are given a sample set at each iteration, and our goal is just to identify a Λ–poised subset. The constant $\Lambda > 0$ is chosen at iteration $k = 0$.

The notion of the simplex gradient can be extended to higher order derivatives [6]. One can consider the computation of a simplex Hessian by extending the linear system $S^\top g = \delta(f;S)$ to the following system in the variables $g \in \mathbb{R}^n$ and $H \in \mathbb{R}^{n\times n}$, with $H = H^\top$:

$$(3.1)\quad (y^i - y^0)^\top g + \frac{1}{2}(y^i - y^0)^\top H(y^i - y^0) \;=\; f(y^i) - f(y^0), \quad i = 1,\ldots,p.$$

The number of points in the sample set $Y = \{y^0, y^1, \ldots, y^p\}$ must be equal to $p+1 = (n+1)(n+2)/2$ if one wants to compute a full symmetric simplex Hessian. Similarly to the linear case, the simplex gradient $g = \nabla_s f(y^0)$ and the simplex Hessian $H = \nabla_s^2 f(y^0)$, computed from system (3.1) with $p+1 = (n+1)(n+2)/2$ points, coincide with the coefficients of the quadratic multivariate polynomial interpolation model associated with Y. The notions of poisedness and Λ–poisedness and the derivation of the error bounds for simplex Hessians in determined and nondetermined cases are reported in [7, 6].

In our application to a pattern search we are interested in using sample sets with a relatively low number of points. One alternative is to consider fewer points than coefficients in the model and to compute solutions in the minimum norm sense. Another option is to choose to approximate only some portions of the simplex Hessian. For instance, if one is given $2n+1$ points one can compute the n components of a simplex gradient and an approximation to the n diagonal terms of a simplex Hessian. The system to be solved in this case is of the form

$$\begin{bmatrix} y^1 - y^0 & \cdots & y^{2n} - y^0 \\ (1/2)(y^1 - y^0).\hat{}2 & \cdots & (1/2)(y^{2n} - y^0).\hat{}2 \end{bmatrix}^\top \begin{bmatrix} g \\ \mathrm{diag}(H) \end{bmatrix} \;=\; \delta(f;S),$$

where $\delta(f;S) = [\, f(y^1) - f(y^0), \ldots, f(y^{2n}) - f(y^0)\,]^\top$ and the notation ".^2" stands for componentwise squaring. Once again, if the number of points is lower than $2n+1$ a minimum norm solution can be computed.

4. Ordering the polling in a pattern search. A pattern search method generates a number of function evaluations at each iteration. One can store some of these points and corresponding objective function values during the course of the iterations. Thus, at the beginning of each iteration, one can try to identify a subset of these points with desirable geometrical properties (Λ–poisedness in our context).

If successful in such an attempt, we compute some form of simplex derivatives, such as a simplex gradient. We can then compute, at no additional cost, a direction of potential descent or of potential steepest descent (a negative simplex gradient, for example). We call such a direction a *descent indicator*. There may be iterations (especially at the beginning) in which we fail to compute a descent indicator, but such failures cost no extra function evaluations either.

Our main goal is to use descent indicators based on simplex derivatives to order the poll vectors efficiently in the poll step. We can also explore the use of simplex derivatives in other components of a pattern search method such as the search step or the mesh size parameter update.

We adapt the description of a pattern search to follow the approach described above. The class of pattern search methods remains essentially the same and is spelled out in Figure 4.1. All modifications to the algorithm reported in Figure 2.1 are marked in italics in Figure 4.1 for better identification.

Pattern Search Method — Using Sampling and Simplex Derivatives

Initialization
Choose x_0 and $\alpha_0 > 0$. Choose a positive spanning set D. Select all constants needed for procedures `[search]`, `[order]`, and `[mesh]`. Set $k = 0$. *Set $X_0 = [x_0]$ to initialize the list of points maintained by* `[store]`. *Choose a maximum number p_{max} of points that can be stored. Choose also the minimum s_{min} and the maximum s_{max} number of points involved in any simplex derivatives calculation ($2 \leq s_{min} \leq s_{max}$). Choose $\Lambda > 0$ and $\sigma_{max} \geq 1$.*

Identifying a Λ–poised sample set and computing simplex derivatives
Skip this step if there are not enough points, i.e., if $|X_k| < s_{min}$. Set $\Delta_k = \sigma_k \alpha_{k-1} \max_{b \in B_{k-1}} \|b\|$, where $\sigma_k \in [1, \sigma_{max}]$. Try to identify a set of points Y_k in $X_k \cap \mathcal{B}(x_k; \Delta_k)$, with as many points as possible (up to s_{max}) and such that Y_k is Λ–poised and includes the current iterate x_k. If $|Y_k| \geq s_{min}$, compute some form of simplex derivatives based on Y_k (and from that compute a descent indicator d_k).

Search step
Call `[search]` to try to compute a point $x \in M_k$ with $f(x) < f(x_k)$ by evaluating the function only at a finite number of points in M_k *and calling* `[store]` *each time a point is evaluated.* If such a point is found, then set $x_{k+1} = x$, declare the iteration as successful, and skip the poll step.

Poll step
Choose a positive basis $B_k \subset D$. Call `[order]` to order the polling set $P_k = \{x_k + \alpha_k b : b \in B_k\}$. Start evaluating f at the polling points following the order determined *and calling* `[store]` *each time a point is evaluated.* If a polling point $x_k + \alpha_k b_k$ is found such that $f(x_k + \alpha_k b_k) < f(x_k)$, then stop polling, set $x_{k+1} = x_k + \alpha_k b_k$, and declare the iteration as successful. Otherwise declare the iteration as unsuccessful and set $x_{k+1} = x_k$.

Updating the mesh size parameter
Call `[mesh]` to compute α_{k+1}. Increment k by one and return to the *simplex derivatives* step.

FIG. 4.1. *The class of pattern search methods used in this paper, adapted now for identifying Λ–poised sample sets and computing simplex derivatives.*

The algorithm maintains a list X_k of evaluated points with maximum size p_{max}. Each time a new point is evaluated, the algorithm calls a new procedure store, which controls the adding (and deleting) of points to X_k.

A new step is included at the beginning of each iteration for computing simplex derivatives. In this step, the algorithm attempts first to extract from X_k a sample set Y_k with the appropriate size and desirable geometrical properties. The points in Y_k must be within a certain distance Δ_k to the current iterate:

$$\Delta_k \ = \ \sigma_k \, \alpha_{k-1} \max_{b \in B_{k-1}} \|b\|,$$

where $\sigma_k \in [1, \sigma_{max}]$, and $\sigma_{max} \geq 1$ is fixed a priori for all iterations. Note that Δ_k is chosen such that $\mathcal{B}(x_k; \Delta_k)$ contains all of the points in $P_{k-1} = \{x_{k-1} + \alpha_{k-1} b : b \in B_{k-1}\}$ when $k-1$ is an unsuccessful iteration. The dependence of Δ_k on α_{k-1} guarantees the asymptotic quality of the simplex derivatives computed at a subsequence of unsuccessful iterates (see Theorems 3.1 and 5.1).

```
procedure order
```
Compute $\cos(d_k, b)$ for all $b \in B_k$. Order the columns in B_k according to decreasing values of the corresponding cosines.

FIG. 4.2. *Ordering the polling vectors according to their angle distance to the descent indicator.*

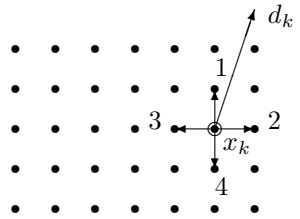


FIG. 4.3. *Ordering the polling vectors using a descent indicator. The positive basis considered is* $B_k = [\, I \; -I \,]$.

We consider two simple strategies for deciding whether or not to store a point, once the function has been evaluated there:

- `store-succ`: keeps only the successful iterates x_{k+1} (for which $f(x_{k+1}) < f(x_k)$);
- `store-all`: keeps every evaluated point.

In both cases, points are added sequentially to X_k at the top of the list. In store-succ, the points in the list X_k are ordered by increasing objective function values. When (and if) X_k has reached its predetermined size p_{max}, we must first remove a point before adding a new one. We assume that the points are removed from the end of the list. Note that both variants store successful iterates x_{k+1} (for which $f(x_{k+1}) < f(x_k)$). Clearly, the current iterate x_k is always in X_k, when store-succ is chosen. However, for store-all, x_k could be removed from the list if a number of consecutive unsuccessful iterates occur. We must therefore add a safeguard to prevent this from happening.

Having a descent indicator d_k at hand, we can order the polling vectors according to increasing magnitudes of the angles between d_k and the polling directions. So the first polling point to be evaluated is the one corresponding to the polling vector making the smallest angle with d_k. We describe this procedure order in Figure 4.2 and illustrate it in Figure 4.3.

The descent indicator could be a negative simplex gradient $d_k = -\nabla_s f(x_k)$, where $S_k = [\, y_k^1 - x_k \;\; \ldots \;\; y_k^{q_k} - x_k \,]$ is formed from the sample set $Y_k = \{y_k^0, y_k^1, \ldots, y_k^{q_k}\}$, with $q_k + 1 = |Y_k|$ and $y_k^0 = x_k$. We designate this approach by sgradient. Another possibility is to compute $d_k = -H_k^{-1} g_k$, where g_k is a simplex gradient and H_k approximates a simplex Hessian. In section 8, we test numerically the diagonal simplex Hessians described at the end of section 3. This approach is designated by shessian.

5. Geometry of the sample sets. If evaluated points are added to the list X_k according to the store-all criterion, it is possible to guarantee the quality of the sample sets Y_k used to compute the simplex derivatives after the occurrence of unsuccessful iterations.

Let us focus on the case where our goal is to compute simplex gradients. We define

$$s_{pb} \;=\; \min\{|B| : \; B \subset D, \; B \text{ positive basis}\}.$$

First we assume that $s_{min} \leq s_{pb}$, i.e., that simplex gradients can be computed from s_{pb} points of X_k with appropriate geometry. If iteration $k-1$ was unsuccessful, then at least $|B_{k-1}|$ points were added to X_{k-1} (the polling points $x_{k-1}+\alpha_{k-1}b$ for all $b \in B_{k-1}$). Such points are part of X_k as well as the current iterate $x_k = x_{k-1}$. It is shown in the next theorem that the sample set $Y_k = \{x_k\} \cup \{x_{k-1}+\alpha_{k-1}b : \ b \in B_{k-1}\} \subset X_k$ is poised for a simplex gradient calculation.

It is also shown that the sample set $Y_k \subset X_k$ formed by x_k and by only $|B_{k-1}|-1$ of the points $x_{k-1}+\alpha_{k-1}b$, $b \in B_{k-1}$, is also poised for a simplex gradient calculation. In this case, we set $s_{min} \leq s_{pb}-1$.

THEOREM 5.1. *Let $k-1$ be any unsuccessful iteration of the pattern search method of Figure* 4.1 *using the store-all strategy.*

- *Suppose $s_{min} \leq s_{pb}$. There exists a positive constant Λ_1 (independent of k) such that the sample set $Y_k \subset X_k$ formed by $x_k = x_{k-1}$ and the points $x_{k-1}+\alpha_{k-1}b$, $b \in B_{k-1}$, is Λ_1–poised for a (overdetermined) simplex gradient calculation.*
- *Suppose $s_{min} \leq s_{pb}-1$. There exists a positive constant Λ_2 (independent of k) such that the sample set $Y_k \subset X_k$ formed by $x_k = x_{k-1}$ and by only $|B_{k-1}|-1$ of the points $x_{k-1}+\alpha_{k-1}b$, $b \in B_{k-1}$, is Λ_2–poised for a (determined or overdetermined) simplex gradient calculation.*

Proof. To simplify the notation we write $B = B_{k-1}$. To prove the first statement, let $Y_k = \{y_k^0, y_k^1, \ldots, y_k^{q_k}\}$, with $q_k+1 = |Y_k| = |B|+1$ and $y_k^0 = x_k$. Then

$$S_k \ = \ [\, y_k^1 - x_k \ \ldots \ y_k^{q_k} - x_k \,] \ = \ [\, \alpha_{k-1}b_1 \ \ldots \ \alpha_{k-1}b_{|B|} \,] \ = \ \alpha_{k-1}B.$$

The matrix B has rank n since it linearly spans $\mathbb{R}^n$ by definition. Thus,

$$\frac{1}{\Delta_k}S_k \ = \ \frac{\alpha_{k-1}}{\sigma_k\,\alpha_{k-1}\max_{b\in B}\|b\|}B \ = \ \frac{1}{\sigma_k}\frac{1}{\max_{b\in B}\|b\|}B,$$

and the geometry constant associated with this sample set Y_k is given by

$$\frac{1}{\sigma_k}\|\Sigma^{-1}\| \qquad \text{with} \qquad \frac{1}{\max_{b\in B}\|b\|}B^\top \ = \ U\Sigma V^\top.$$

Since $\sigma_k \geq 1$, if we choose the poisedness constant such that

$$\Lambda_1 \ \geq \ \max\left\{\|\Sigma^{-1}\| : \ \begin{array}{l} \frac{1}{\max_{b\in B}\|b\|}B^\top \ = \ U\Sigma V^\top \\ \forall\, \text{positive bases } B \subset D \end{array}\right\},$$

then we are guaranteed to identify a Λ_1–poised sample set after any unsuccessful iteration.

In the second case, we have $q_k+1 = |Y_k| = |B|$ and

$$S_k \ = \ \alpha_{k-1}B_{|B|-1},$$

where $B_{|B|-1}$ is some column submatrix of B with $|B|-1$ columns. Since B is a positive spanning set, $B_{|B|-1}$ linearly spans $\mathbb{R}^n$ (see [10, Theorem 3.7]), and therefore it has rank n. The difference now is that we must consider all submatrices $B_{|B|-1}$ of B. Thus, if we choose the poisedness constant such that

$$\Lambda_2 \ \geq \ \max\left\{\|\Sigma^{-1}\| : \ \begin{array}{l} \frac{1}{\max_{b\in B}\|b\|}B_{|B|-1}^\top \ = \ U\Sigma V^\top \\ \forall\ B_{|B|-1} \subset B,\ \forall\, \text{positive bases } B \subset D \end{array}\right\},$$

we are guaranteed to identify a Λ_2–poised sample set after any unsuccessful iteration. ∎

We point out that a result of this type is not necessarily restricted to unsuccessful iterations. Other geometry scenarios can be explored at successful iterations.

6. Pruning the polling directions. Abramson, Audet, and Dennis [1] show that, for a special choice of the positive spanning set D, rough approximations to the gradient of the objective function can be used to reduce the polling step to a single function evaluation. The gradient approximations considered were ϵ–approximations to the large components of the gradient vector.

Let g be a nonzero vector in $\mathbb{R}^n$ and $\epsilon \geq 0$. Consider

$$J^\epsilon(g) \;=\; \{i \in \{1,\ldots,n\} : |g_i| + \epsilon \geq \|g\|_\infty\}\,,$$

and for every $i \in \{1,\ldots,n\}$ let

$$d^\epsilon(g)_i \;=\; \begin{cases} \operatorname{sign}(g_i) & \text{if } i \in J^\epsilon(g), \\ 0 & \text{otherwise.} \end{cases} \tag{6.1}$$

The vector g is said to be an ϵ–approximation to the large components of a nonzero vector $v \in \mathbb{R}^n$ if and only if $i \in J^\epsilon(g)$ whenever $|v_i| = \|v\|_\infty$ and $\operatorname{sign}(g_i) = \operatorname{sign}(v_i)$ for every $i \in J^\epsilon(g)$.

The question that arises now is whether a descent indicator d_k, and, in particular, a negative simplex gradient $-\nabla_s f(x_k)$, is an ϵ–approximation to the large components of $-\nabla f(x_k)$ for some $\epsilon > 0$. We show in the next theorem that the answer is affirmative, provided that the mesh size parameter α_k is sufficiently small, an issue we readdress at the end of this section.

We will use the notation previously introduced in this paper. We consider a sample set Y_k and the corresponding matrix S_k. The set Y_k is included in the ball $\mathcal{B}(x_k;\Delta_k)$ centered at x_k with radius $\Delta_k = \sigma_k\,\alpha_{k-1}\,\max_{b\in B_{k-1}}\|b\|$, where B_{k-1} is the positive basis used for polling at the previous iteration.

THEOREM 6.1. *Let Y_k be a Λ–poised sample set (for simplex gradients) computed at iteration k of a pattern search method, with $q_k + 1 \geq n + 1$ points.*

Assume that ∇f is Lipschitz continuous in an open domain Ω containing $\mathcal{B}(x_k;\Delta_k)$ with constant $\gamma > 0$.

Then, if

$$\alpha_k \;\leq\; \frac{\|\nabla f(x_k)\|_\infty}{\sqrt{q_k}\gamma\Lambda\sigma_{max}\max_{b\in B_{k-1}}\|b\|}, \tag{6.2}$$

the negative simplex gradient $-\nabla_s f(x_k)$ is an ϵ_k–approximation to the large components of $-\nabla f(x_k)$, where

$$\epsilon_k \;=\; \left(q_k^{\frac{1}{2}}\gamma\Lambda\sigma_{max}\max_{b\in B_{k-1}}\|b\|\right)\alpha_k.$$

Proof. For i in the index set

$$I_k \;=\; \{i \in \{1,\ldots,n\} :\; |\nabla f(x_k)_i| = \|\nabla f(x_k)\|_\infty\},$$

we get from Theorem 3.1 that

$$\begin{aligned} \|\nabla_s f(x_k)\|_\infty &\leq \|\nabla f(x_k) - \nabla_s f(x_k)\|_\infty + |\nabla f(x_k)_i| \\ &\leq 2\|\nabla f(x_k) - \nabla_s f(x_k)\| + |\nabla_s f(x_k)_i| \\ &\leq q_k^{\frac{1}{2}}\gamma\Lambda\Delta_k + |\nabla_s f(x_k)_i| \\ &\leq \epsilon_k + |\nabla_s f(x_k)_i|. \end{aligned}$$

From Theorem 3.1 we also know that

$$-\nabla_s f(x_k)_i \;=\; -\nabla f(x_k)_i + \xi_{k,i}, \quad \text{where} \quad |\xi_{k,i}| \le q_k^{\frac{1}{2}} \frac{\gamma}{2} \Lambda \Delta_k.$$

If $-\nabla f(x_k)_i$ and $\xi_{k,i}$ are equally signed, so are $-\nabla f(x_k)_i$ and $-\nabla_s f(x_k)_i$. Otherwise, they are equally signed if

$$|\xi_{k,i}| \;\le\; q_k^{\frac{1}{2}} \frac{\gamma}{2} \Lambda \Delta_k \;\le\; \frac{1}{2} \|\nabla f(x_k)\|_\infty \;=\; \frac{1}{2} |\nabla f(x_k)_i|.$$

The proof is concluded using the expression for Δ_k and the bound for α_k given in the statement of the theorem. □

Theorem 4 by Abramson, Audet, and Dennis [1] shows that an ϵ–approximation prunes the set of the polling directions to a singleton, when considering

$$D \;=\; \{-1, 0, 1\}^n$$

and the positive spanning set

$$D_k \;=\; \{d^\epsilon(g_k)\} \cup \mathbb{A}(-\nabla f(x_k)),$$

where g_k is an ϵ–approximation to $-\nabla f(x_k)$, $d^\epsilon(\cdot)$ is defined in (6.1), and

$$\mathbb{A}(-\nabla f(x_k)) \;=\; \{d \in D : -\nabla f(x_k)^\top d < 0\}$$

represents the set of the ascent directions in D. The pruning is to the singleton $\{d^\epsilon(g_k)\}$, meaning that $d^\epsilon(g_k)$ is the only vector d in D_k such that $-\nabla f(x_k)^\top d \ge 0$.

So, under the hypotheses of Theorem 6.1, it follows that the negative simplex gradient $-\nabla_s f(x_k)$ prunes the positive spanning set

$$D_k \;=\; \{d^{\epsilon_k}(-\nabla_s f(x_k))\} \cup \mathbb{A}(-\nabla f(x_k))$$

to a singleton, namely, $\{d^{\epsilon_k}(-\nabla_s f(x_k))\}$, where ϵ_k is given in Theorem 6.1.

Now we analyze in more detail the role of condition (6.2). There is no guarantee that this condition on α_k can be satisfied asymptotically. Condition (6.2) gives us only an indication of the pruning effect of the negative simplex gradient, and it is more likely to be satisfied at points where the gradient is relatively large. What is known is actually a condition that shows that α_k dominates $\|\nabla f(x_k)\|$ at unsuccessful iterations k:

$$\|\nabla f(x_k)\| \;\le\; \left(\gamma \kappa(B_k)^{-1} \max_{b \in B_k} \|b\|\right) \alpha_k,$$

where

$$\kappa(B_k) \;=\; \min_{d \in \mathbb{R}^n; d \neq 0} \max_{b \in B_k} \frac{d^\top b}{\|d\| \|b\|} \;>\; 0$$

is the cosine measure of the positive basis B_k (see [17, Theorem 3.3]). Since only a finite number of positive bases is used, $\kappa(B_k)^{-1}$ is uniformly bounded. So one can be assured that at unsuccessful iterations the norm of the gradient is bounded by a constant times α_k.

However, it has been observed in [11] that, for some problems, α_k goes to zero faster than $\|\nabla f(x_k)\|$. Our numerical experience with pattern search has also pointed us in this direction. It is more difficult, however, to sharply verify condition (6.2), since it depends on the Lipschitz constant of ∇f. A detailed numerical study of these asymptotic behaviors is beyond the scope of this paper.

procedure mesh-sd

The constants τ and ξ must satisfy $\tau \in \mathbb{Q}$, $\tau > 1$, and $\xi > 0$ and should be initialized at iteration $k = 0$ together with $j_{max} \in \mathbb{Z}$, $j_{max} \geq 0$, and $j_{min} \in \mathbb{Z}$, $j_{min} \leq -1$. The exponents satisfy $j_k^+ \in \{0, 1, 2, \ldots, j_{max}\}$ and $j_k^- \in \{j_{min}, \ldots, -1\}$.

If the iteration was successful, then compute

$$\rho_k \;=\; \frac{f(x_k) - f(x_{k+1})}{m_k(x_k) - m_k(x_{k+1})}.$$

If $\rho_k > \xi$, then $\alpha_{k+1} = \tau^{j_k^+} \alpha_k$,
If $\rho_k \leq \xi$, then $\alpha_{k+1} = \alpha_k$.

If the iteration was unsuccessful, then contract mesh by decreasing the mesh size parameter $\alpha_{k+1} = \tau^{j_k^-} \alpha_k$.

FIG. 7.1. *Updating the mesh size parameter (using a sufficient decrease but meeting rational lattice requirements).*

7. Other uses for simplex derivatives. Having computed before some form of simplex derivatives, one can use the available information for purposes other than ordering the polling vectors. In this section, we suggest two other uses for simplex derivatives in pattern search: the update of the mesh size parameter and the computation of a search step.

When a simplex gradient $\nabla_s f(x_k)$ is computed, a linear model $m_k(y) = f(x_k) + \nabla_s f(x_k)^\top (y - x_k)$ can be used to update the mesh size parameter α_k by imposing a sufficient decrease condition. In this case, we set

$$\rho_k \;=\; \frac{f(x_k) - f(x_{k+1})}{m_k(x_k) - m_k(x_{k+1})} \;=\; \frac{f(x_k) - f(x_{k+1})}{-\nabla_s f(x_k)^\top (x_{k+1} - x_k)}.$$

If x_{k+1} is computed in a successful poll step, then $x_{k+1} - x_k = \alpha_k b_k$ for some $b_k \in B_k$. In the quadratic case, the model is replaced by $m_k(y) = f(x_k) + g_k^\top (y - x_k) + (1/2)(y - x_k)^\top H_k (y - x_k)$. We call this procedure mesh-sd and describe it in Figure 7.1, where the sufficient decrease is applied only to successful iterations.

Since the expansion and contraction parameters are restricted to integer powers of τ and since the contraction rules match what was given in the mesh procedure of Figure 2.2, the modification introduced in mesh-sd has no influence on the global convergence properties of the underlying pattern search method.

There are many possibilities for a search step. One possibility is to first form a surrogate model $m_k(y)$ based on some form of simplex derivatives computed using the sample set Y_k and then to minimize this model in $\mathcal{B}(x_k; \Delta_k)$, after which we would project the minimizer onto the mesh M_k. We described above two examples of such a model $m_k(y)$, but many others could be considered. The use of surrogate models in the search step is the topic of separate research.

8. Implementation and numerical results. To serve as a baseline for numerical comparisons, we have implemented a basic pattern search algorithm of the form given in Figure 2.1. Specifically, no search step is used, the mesh size parameter is left unchanged at successful iterations, and points in the poll step are always evaluated in the same consecutive order as originally stored. We refer to this version of pattern search as basic.

TABLE 8.1
Test set and results for the basic version.

		Positive basis			
		$D = [-e\ I]$		$D = [I\ -I]$	
Problem	Dimension	fevals	fvalue	fevals	fvalue
arwhead	10	1068	4.19e-09	361	0.00e+00
arwhead	20	3718	8.85e-09	721	0.00e+00
bdqrtic	10	2561	1.19e+01	948	1.19e+01
bdqrtic	20	19038	3.54e+01	4120	3.54e+01
bdvalue	10	36820	4.39e-07	33077	4.39e-07
bdvalue	20	255857	1.30e-05	245305	1.29e-05
biggs6	6	339840	6.50e-03	467886	9.58e-06
brownal	10	468150	1.84e+00	74922	2.02e-06
brownal	20	1073871	1.55e+01	284734	1.04e-05
broydn3d	10	2281	3.26e-08	1743	4.52e-09
broydn3d	20	17759	2.91e-07	6868	2.47e-08
integreq	10	2595	4.42e-09	1034	2.35e-10
integreq	20	20941	3.20e-08	4244	4.86e-10
penalty1	10	552357	7.33e-05	234274	7.09e-05
penalty1	20	999305	1.66e-04	535100	1.58e-04
penalty2	10	46696	4.09e-04	496275	4.04e-04
penalty2	20	366131	8.32e-03	1494751	8.30e-03
powellsg	12	192270	1.85e-04	58987	9.85e-07
powellsg	20	480158	3.08e-04	158591	1.64e-06
srosenbr	10	401321	6.83e-05	171061	6.83e-05
srosenbr	20	1076983	2.68e-02	649621	1.37e-04
tridia	10	1000805	5.95e-01	901720	5.85e-01
tridia	20	20483	6.24e-01	6635	6.24e-01
vardim	10	251599	2.23e-05	86316	6.64e-07
vardim	20	961697	1.76e+04	1230761	8.71e-04
woods	12	164675	1.02e-04	110662	3.78e-05
woods	20	435786	3.53e-04	300296	6.29e-05

We have tested a number of pattern search methods of the form described in Figure 4.1. The strategies order (Figure 4.2) and mesh-sd (Figure 7.1) were run in four different modes according to the way of storing points (store-succ or store-all) and to the way of computing simplex derivatives and descent indicators (sgradient or shessian). Moreover, we implemented the strategy suggested in [14] and described in section 2 for updating the mesh size parameter (here named as mesh-HKT) and the dynamic polling strategy suggested in [4] for changing the order of the polling directions (see section 2). We tested a very crude search step based on taking a step along the descent indicator with a step size of the order of α_k (see [9] for the details).

The algorithms were coded in MATLAB and ran on 27 unconstrained problems belonging to the CUTEr collection [13], gathered mainly from papers on derivative-free optimization. The objective functions of these problems are twice continuously differentiable. Their dimensions are given in Table 8.1. The starting points used were those reported in CUTEr. Problems bdvalue, integreq, and broydn3d were posed as unconstrained optimization problems like originally in [19]. The stopping criterion consisted of the mesh size parameter becoming lower than 10^{-5} or a maximum number of 100000 iterations being reached.

The simplex derivatives were computed based on Λ-poised sets Y_k, where $\Lambda = 100$. The factor σ_k was chosen as 1 ($k-1$ unsuccessful), 2 ($k-1$ successful and $\alpha_k = \alpha_{k-1}$), and 4 ($k-1$ successful and $\alpha_k > \alpha_{k-1}$). The values for the parameters s_{min}, s_{max}, and p_{max} are given in Table 8.2. We started all runs with the mesh size parameter $\alpha_0 = 1$. In all versions, the contraction factor was set to $\tau^{j_k^-} = 0.5$, and the expansion

TABLE 8.2
Sizes of the list X_k and of the set Y_k.

	sgradient		shessian	
Size	store-succ	store-all	store-succ	store-all
p_{max}	$2(n+1)$	$4(n+1)$	$4(n+1)$	$8(n+1)$
s_{min}	$(n+1)/2$	$n+1$	n	$2n+1$
s_{max}	$n+1$	$n+1$	$2n+1$	$2n+1$

factor (when used) was set to $\tau^{j_k^+} = 2$. In the mesh-sd strategy of Figure 7.1, we set ξ equal to 0.75.

We draw conclusions based on two positive bases: $[\,I\ -I\,]$ and $[\,-e\ I\,]$. The maximal positive basis $[\,I\ -I\,]$ corresponds to a coordinate search, and it provided the best results for the basic version among a few positive bases stored in different orders (which included $[\,I\ -I\,]$, $[\,-I\ I\,]$, $[\,-e\ I\,]$, $[\,e\ -I\,]$, $[\,I\ -e\,]$, $[\,-I\ e\,]$, and a minimal basis with angles between vectors of uniform amplitude). The positive basis stored as $[\,-e\ I\,]$ was the minimal positive basis which behaved the best. In Table 8.1 we report the results obtained by the basic version for these two positive bases.

By combining all possibilities, we tested a total of 120 versions, 112 involving simplex derivatives. A summary of the complete numerical results is reported in [9].

8.1. Discussion based on complete results. First, we point out that 91% of the versions involving simplex derivatives lead to an average decrease in the number of function evaluations [9]. Moreover, 61 out of the 112 strategies tested provided a negative 75% percentile for the variation in the number of function evaluations. This means that for each of these 61 strategies, a reduction in the number of function evaluations was achieved for 75% of the problems tested.

The overall results [9] showed a superiority of sgradient over shessian, which is not surprising because the number of points required to identify Λ–poised sets in sgradient is lower than in shessian. Also, another reason for sgradient being possibly better than shessian is that, if the simplex gradient is sufficiently close to the true gradient, then directions making a small angle with the negative simplex gradient will be descent directions, while the same is not guaranteed when we use simplex Newton directions. Some shessian versions, however, have behaved relatively well [9].

For the positive basis $[\,I\ -I\,]$ there is a clear gain when using store-all compared to store-succ [9]. However, for the positive basis $[\,-e\ I\,]$, the advantage of store-all over store-succ is not as clear [9]. In general, the advantage of store-all may be explained by the frequent number of unsuccessful iterations that tend to occur in the last iterations of a pattern search run. The effect of the poll ordering is also more visible when using the positive basis $[\,I\ -I\,]$, due to the larger number of polling vectors.

Strategies mesh-sd and mesh-HKT made a clear positive impact when using the smaller positive basis $[\,-e\ I\,]$ (see [9]). This effect was lost in the larger positive basis $[\,I\ -I\,]$, where the order procedure seems to perform well on its own for this test set.

8.2. Discussion based on best results. We report in Table 8.3 a summary of the results for a number of versions based on $[\,I\ -I\,]$. Included in this restricted set of versions are the ones that lead to the best results among all of the 120 versions tested. (The results for the remaining versions are summarized in [9].)

An explanation about Table 8.3 is in order. For each strategy and for each problem, we calculated the percentage of iterations that used simplex descent indicators as well as the variation in the number of function evaluations required relatively to the basic version. These percentages were grouped by strategy, and their average values are

TABLE 8.3

Average percentage of iterations that used simplex descent indicators (second column), average variation of function evaluations by comparison to the basic version (third column), and cumulative percentages for the optimal gaps of the final iterates (fourth to sixth columns). Case sgradient and $D = [\, I \; -I \,]$.

Strategy	% poised	Number of evaluations	Optimal gap		
			10^{-7}	10^{-4}	10^{-1}
`basic`	—	—	33.33%	81.48%	92.59%
`mesh-HKT`	—	+4.02%	40.74%	81.48%	92.59%
`dynamic polling`	—	-10.99%	33.33%	81.48%	92.59%
`mesh-HKT,dynamic polling`	—	-15.17%	48.15%	81.48%	92.59%
`mesh-sd (store-succ)`	14.40%	-3.07%	33.33%	81.48%	92.59%
`mesh-sd (store-all)`	73.33%	+0.45%	33.33%	81.48%	92.59%
`order (store-all)`	27.26%	-51.16%	37.04%	85.19%	92.59%
`mesh-sd,order (store-all)`	28.56%	-51.47%	37.04%	85.19%	92.59%
`mesh-HKT,order (store-all)`	58.51%	-54.22%	51.85%	81.48%	88.89%

reported in the second and third columns of Table 8.3. The last three columns of the table represent the cumulative percentages for the optimal gaps of the final iterates.

The quality of the final objective function values obtained for the versions included in Table 8.3 is comparable to the basic version, as one see from the final cumulative optimal gaps reported.

It is clear that none of the strategies for updating the step size parameter (mesh-sd and mesh-HKT) made improvements on their one, the former being slightly better than the latter.

The best result without using simplex derivatives was obtained by combining dynamic polling and mesh-HKT (15% less function evaluations than the basic version).

Three versions that incorporated order reached a reduction of around 50% in the number of function evaluations. The order procedure in the store-all mode leads, on its own, to a 51% improvement, compared to 11% of dynamic polling.

The best version achieved a reduction of 54% in the number of evaluations by combining order and mesh-HKT in the store-all mode. In Table 8.4 we report the results obtained by this version as well as by the version that applies only the order procedure in the store-all mode.

8.3. Additional tests. We picked some of these problems and ran several versions for $n = 40$ and $n = 80$. Our conclusions remain essentially the same. The ratios of improvement in the number of function evaluations and the quality of the final iterates do not change significantly with the increase of the dimension of the problem but rather with the increase of the number of polling vectors in the positive spanning set or with the increase in its cosine measure (both of which happen, for instance, when going from $[\, -e \; I \,]$ to $[\, I \; -I \,]$).

We also tried to investigate how sensitive the different algorithmic versions are to the choice of the parameter ξ used in the mesh-sd strategy. We tried other values (for instance, 0.5 and 0.95), but the results did not improve.

We repeated these computational tests on a different set of test problems, consisting of seven randomly generated quadratic functions, each one of dimension 10. The quadratic functions were defined by $f(x) = x^\top Ax$, where $A = B^\top B$ and B is a matrix with random normal entries of mean 0 and standard deviation 1. We also randomly generated the starting point for the algorithm, using the same normal distribution. Once again, the conclusions for the different strategies remained essentially the same

TABLE 8.4
Results for the best versions, using the positive basis $D = [\, I \; -I \,]$.

		Strategy			
		`order`		`order,mesh-HKT`	
Problem	Dimension	fevals	fvalue	fevals	fvalue
arwhead	10	361	0.00e+00	361	0.00e+00
arwhead	20	721	0.00e+00	721	0.00e+00
bdqrtic	10	696	1.19e+01	696	1.19e+01
bdqrtic	20	2138	3.54e+01	2138	3.54e+01
bdvalue	10	34922	6.89e-07	28411	6.52e-07
bdvalue	20	255989	1.66e-05	213297	1.65e-05
biggs6	6	105592	4.50e-07	164168	4.53e-07
brownal	10	21045	1.88e-06	38398	2.44e-06
brownal	20	67152	6.16e-06	4227	1.00e+00
broydn3d	10	917	4.73e-09	917	4.73e-09
broydn3d	20	2940	2.35e-08	2940	2.35e-08
integreq	10	597	2.35e-10	597	2.35e-10
integreq	20	1573	4.86e-10	1573	4.86e-10
penalty1	10	126307	7.09e-05	177360	7.09e-05
penalty1	20	229825	1.58e-04	279491	1.58e-04
penalty2	10	55087	4.04e-04	93192	4.05e-04
penalty2	20	189446	8.29e-03	355154	8.29e-03
powellsg	12	594	0.00e+00	614	0.00e+00
powellsg	20	45258	1.31e-06	8702	2.81e-11
srosenbr	10	136327	6.83e-05	119830	6.83e-05
srosenbr	20	567937	1.37e-04	358656	1.36e-04
tridia	10	539119	5.85e-01	908097	5.89e-01
tridia	20	2724	6.24e-01	2828	6.24e-01
vardim	10	5382	2.29e-07	6550	9.15e-08
vardim	20	67487	9.27e-06	71692	1.68e-06
woods	12	59565	3.94e-05	577	0.00e+00
woods	20	106339	6.55e-05	1064	0.00e+00

(with improvement of the results for the minimal positive basis $[\, -e \; I \,]$). We used these examples to study the descent properties of the negative simplex gradient. In our experiments, the simplex gradient made an acute angle with the true gradient on average in 77% of the cases where it was computed. These occurrences tend to happen more towards the end of the runs when the mesh size parameter gets smaller.

8.4. Pruning. To better understand the theoretical results derived in section 6, we implemented a computational variant of the pruning strategy. We did not consider the generating set $D = \{-1, 0, 1\}^n$, as suggested by Abramson, Audet, and Dennis [1], nor did we verify condition (6.2) (or some approximated form of it by estimating the Lipschitz constant involved) before pruning the polling vectors. As a result, we are violating the conditions required for the analysis of the pruning strategy. We tested two different variants for pruning the positive bases $[\, I \; -I \,]$ and $[\, -e \; I \,]$: (i) pruning to a single direction, namely, the one that makes the angle of smallest amplitude with the descent indicator and (ii) pruning to all of the directions that make an acute angle with the descent indicator.

To reach a final iterate of quality nearly similar to the one obtained by the basic version, we had to use the positive basis $[\, I \; -I \,]$ and prune with more than one direction. In this case, pruning achieved an average reduction in the number of function evaluations of 10% and 42%, for the store-succ and store-all variants, respectively. Pruning tends to generate less polling points, which in turn decreases the chances of building well-poised sets.

More research is needed in order to evaluate the potential of the negative simplex gradient as an ϵ-approximation to the large components of the negative gradient vector and its use for pruning the polling directions. The use of the generating set $D = \{-1, 0, 1\}^n$ and the implementation of some form of the condition (6.2) might have a positive impact.

9. Concluding remarks and future work. We have proposed the use of simplex derivatives in pattern search methods in two ways: ordering the polling vectors and updating the mesh size parameter. For the calculation of the simplex derivatives, we considered sample sets constructed in two variants: storing only all recent successful iterates or storing all recent points where the objective function was evaluated. Finally, we studied two types of simplex derivatives: simplex gradients and diagonal simplex Hessians. It is important to remark that the incorporation of these strategies in a pattern search is done at no further expense in function evaluations.

The introduction of simplex derivatives in pattern search methods can lead to a significant reduction in the number of function evaluations for the same quality of the final iterates.

As a descent indicator, we recommend the use of the negative simplex gradient over the simplex Newton direction. In fact, most of the iterations of a pattern search run are performed for small values of the mesh size parameter. In such cases, the negative gradient is better than the Newton direction as an indicator for descent, and the same argument applies to their simplex counterparts.

For a coordinate search ($D = [\, I \; -I \,]$), ordering the polling directions according to a simplex descent indicator (negative simplex gradient) made a significant impact in the reduction of the number of function evaluations. For this type of positive basis, storing all recent points where the objective function was evaluated seems to be the best approach.

Our numerical findings showed that updating the mesh size parameter based on a sufficient decrease condition can be worthwhile applying when using minimal positive bases (such as $D = [\, -e \; I \,]$). In such cases, storing only all recent successful iterates may also be advantageous.

There are at least two natural generalizations of the ideas presented in this paper. One is to apply simplex derivative-based strategies to improve parallel versions of a pattern search. Another generalization consists of analyzing the properties of simplex gradients when direct search methods are applied to nonsmooth functions [8]. The use of simplex derivatives in the design of an efficient search step is also the subject of future research.

Acknowledgments. During the course of the revision of this paper, the first author visited the College of William & Mary. The authors are grateful to Robert Michael Lewis, Virginia Torczon, and Michael Trosset for many interesting and stimulating comments, which have contributed to improvements in the manuscript. We also thank the anonymous referees for their suggestions.

REFERENCES

[1] M. A. Abramson, C. Audet, and J. E. Dennis, jr., *Generalized pattern searches with derivative information*, Math. Program., 100 (2004), pp. 3–25.

[2] P. Alberto, F. Nogueira, H. Rocha, and L. N. Vicente, *Pattern search methods for user-provided points: Application to molecular geometry problems*, SIAM J. Optim., 14 (2004), pp. 1216–1236.

[3] C. Audet and J. E. Dennis, jr., *Analysis of generalized pattern searches*, SIAM J. Optim., 13 (2002), pp. 889–903.
[4] C. Audet and J. E. Dennis, jr., *Mesh adaptive direct search algorithms for constrained optimization*, SIAM J. Optim., 17 (2006), pp. 188–217.
[5] D. M. Bortz and C. T. Kelley, *The simplex gradient and noisy optimization problems*, in Computational Methods in Optimal Design and Control, Progr. Syst. Control Theory 24, J. T. Borggaard, J. Burns, E. Cliff, and S. Schreck, eds., Birkhäuser, Boston, 1998, pp. 77–90.
[6] A. R. Conn, K. Scheinberg, and L. N. Vicente, *Geometry of Sample Sets in Derivative Free Optimization: Polynomial Regression and Underdetermined Interpolation*, Technical report 05-15, Departamento de Matemática, Universidade de Coimbra, Portugal, 2005.
[7] A. R. Conn, K. Scheinberg, and L. N. Vicente, *Geometry of interpolation sets in derivative free optimization*, Math. Program., to appear.
[8] A. L. Custódio, J. E. Dennis, jr., and L. N. Vicente, *Using Simplex Gradients of Nonsmooth Functions in Direct Search Methods*, Technical report 06-48, Departamento de Matemática, Universidade de Coimbra, Portugal, 2006.
[9] A. L. Custódio and L. N. Vicente, *Using Sampling and Simplex Derivatives in Pattern Search Methods (Complete Numerical Results)*. See http://www.mat.uc.pt/∼lnv/papers/sid-psm-complete.pdf, 2006.
[10] C. Davis, *Theory of positive linear dependence*, Amer. J. Math., 76 (1954), pp. 733–746.
[11] E. D. Dolan, R. M. Lewis, and V. Torczon, *On the local convergence of pattern search*, SIAM J. Optim., 14 (2003), pp. 567–583.
[12] L. Frimannslund and T. Steihaug, *A generating set search method using curvature information*, Comput. Optim. Appl., to appear.
[13] N. I. M. Gould, D. Orban, and Ph. L. Toint, *CUTEr and SifDec: A constrained and unconstrained testing environment, revisited*, ACM Trans. Math. Software, 29 (2003), pp. 373–394.
[14] P. D. Hough, T. G. Kolda, and V. J. Torczon, *Asynchronous parallel pattern search for nonlinear optimization*, SIAM J. Sci. Comput., 23 (2001), pp. 134–156.
[15] C. T. Kelley, *Detection and remediation of stagnation in the Nelder-Mead algorithm using a sufficient decrease condition*, SIAM J. Optim., 10 (1999), pp. 43–55.
[16] C. T. Kelley, *Iterative Methods for Optimization*, SIAM Front. Appl. Math. 18, SIAM, Philadelphia, 1999.
[17] T. G. Kolda, R. M. Lewis, and V. Torczon, *Optimization by direct search: New perspectives on some classical and modern methods*, SIAM Rev., 45 (2003), pp. 385–482.
[18] R. M. Lewis and V. Torczon, *Rank Ordering and Positive Bases in Pattern Search Algorithms*, Technical report 96-71, ICASE, NASA Langley Research Center, USA, 1996.
[19] J. J. Moré, B. S. Garbow, and K. E. Hillstrom, *Testing unconstrained optimization software*, ACM Trans. Math. Software, 7 (1981), pp. 17–41.
[20] C. J. Price and Ph. L. Toint, *Exploiting problem structure in pattern-search methods for unconstrained optimization*, Optim. Methods Softw., 21 (2006), pp. 479–491.
[21] V. Torczon, *On the convergence of pattern search algorithms*, SIAM J. Optim., 7 (1997), pp. 1–25.
[22] P. Tseng, *Fortified-descent simplicial search method: A general approach*, SIAM J. Optim., 10 (1999), pp. 269–288.

SIAM J. OPTIM.
Vol. 18, No. 2, pp. 556–572

CLARKE SUBGRADIENTS OF STRATIFIABLE FUNCTIONS*

JÉRÔME BOLTE†, ARIS DANIILIDIS‡, ADRIAN LEWIS§, AND MASAHIRO SHIOTA¶

Abstract. We establish the following result: If the graph of a lower semicontinuous real-extended-valued function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ admits a Whitney stratification (so in particular if f is a semialgebraic function), then the norm of the gradient of f at $x \in \operatorname{dom} f$ relative to the stratum containing x bounds from below all norms of Clarke subgradients of f at x. As a consequence, we obtain a Morse–Sard type of theorem as well as a nonsmooth extension of the Kurdyka–Łojasiewicz inequality for functions definable in an arbitrary o-minimal structure. It is worthwhile pointing out that, even in a smooth setting, this last result generalizes the one given in [K. Kurdyka, Ann. Inst. Fourier (Grenoble), 48 (1998), pp. 769–783] by removing the boundedness assumption on the domain of the function.

Key words. Clarke subgradient, Łojasiewicz inequality, critical point, nonsmooth analysis, Whitney stratification

AMS subject classifications. Primary, 49J52; Secondary, 26D10, 32B20

DOI. 10.1137/060670080

1. Introduction.

Nonsmoothness in optimization seldom occurs in an arbitrary manner, but instead is often well-structured. Such structure can often be exploited in sensitivity analysis and algorithm convergence: Examples include "amenability," "subsmoothness," "prox-regularity" (see [32], for example), and more recently the idea of a "partly smooth" function, where a naturally arising manifold $\mathcal{M}$ contains the minimizer and the function is smooth along this manifold. We quote [24] for formal definitions, examples, and more details. In the past two decades, several researchers have tried to capture this intuitive idea in order to develop algorithms ensuring better convergence results: See, for instance, the pioneer work [23] and also [26], [9] for recent surveys.

In this work we shall be interested in a particular class of well-structured (nonsmooth) functions, namely, functions admitting a Whitney stratification (see section 2 for definitions). Since this class contains in particular the semialgebraic and the subanalytic functions (more generally, functions that are definable in some o-minimal structure over $\mathbb{R}$), the derived results can directly be applied in several concrete optimization problems involving such structures. Our central idea is to relate derivative ideas from two distinct mathematical sources: Variational analysis and differential ge-

*Received by the editors September 18, 2006; accepted for publication (in revised form) January 21, 2007; published electronically May 29, 2007.
http://www.siam.org/journals/siopt/18-2/67008.html

†Equipe Combinatoire et Optimisation (UMR 7090), Case 189, Université Pierre et Marie Curie, 4 Place Jussieu, 75252 Paris Cedex 05, France (bolte@math.jussieu.fr, http://www.ecp6.jussieu.fr/pageperso/bolte/). The research of this author was supported by the CRM, Nagoya University and ANR grant ANR-05-BLAN-60248-01 (France).

‡Departament de Matemàtiques, C1/320, Universitat Autònoma de Barcelona, E-08193 Bellaterra (Cerdanyola del Vallès), Spain (arisd@mat.uab.es, http://mat.uab.es/~arisd). The research of this author was supported by MEC grant MTM2005-08572-C03-03 (Spain) and by ANR grant ANR-05-BLAN-60248-01 (France).

§School of Operations Research and Industrial Engineering, Cornell University, 234 Rhodes Hall, Ithaca, NY 14853 (aslewis@orie.cornell.edu, http://www.orie.cornell.edu/~aslewis). The research of this author was supported in part by National Science Foundation grant DMS-0504032.

¶Department of Mathematics, Nagoya University (Furocho, Chikusa), Nagoya 464-8602, Japan (shiota@math.nagoya-u.ac.jp).

ometry. Specifically, we derive a lower bound on the norms of Clarke subgradients at a given point in terms of the "Riemannian" gradient with respect to the stratum containing that point. This is a direct consequence of the "projection formula" given in Proposition 4 and has as corollaries a Morse–Sard type of theorem for Clarke critical points of lower semicontinuous Whitney stratifiable functions (Corollary 5(ii)) as well as a *global* nonsmooth version of the Kurdyka–Łojasiewicz inequality—which is hereby extended to unbounded domains; see Theorem 11—for lower semicontinuous definable functions (Theorem 14 and Corollary 15). Although these results seem natural, analogous ones fail for the (broader) convex-stable subdifferential (introduced and studied in [4]), unless f is assumed to be locally Lipschitz continuous; see Remark 8 and [3].

Theorems of the Morse–Sard type are central in many areas of analysis, typically describing the size of the set of ill-posed problem instances in a given class. Classical results deal with smooth functions [33], [22], but recent advances deal with a variety of nonsmooth settings [3], [13], [14], [15].

A further long-term motivation of this work is to understand the convergence of minimization algorithms. As one example, in order to treat nonconvex (and nonsmooth) minimization problems, the authors of [4] introduced an algorithm called the "gradient sampling algorithm." The idea behind this algorithm was to sample gradients of nearby points of the current iterate and to produce the next iterate by following the vector of minimum norm in the convex hull generated by the sampled negative gradients. In the case that the function is locally Lipschitz, the above method can be viewed as a kind of ε-Clarke subgradient algorithm for which both theoretical and numerical results are quite satisfactory; see [4]. The convergence of the whole sequence of iterates remains, however, an open question, and this is also the case for many classical subgradient methods for nonconvex minimization; see [19]. We hope that, just as in the smooth case, the nonsmooth Łojasiewicz inequality we develop (cf. (22) in section 4) may help in understanding the global convergence of subgradient methods.

As we outline above, we use a stratification approach to develop our results. Ioffe [14] has recently announced an extension of the work described here, leading to a remarkable and powerful Sard-type result for stratifiable multifunctions (see [15]).

2. Preliminaries. In this section we recall several definitions and results concerning nonsmooth analysis (subgradients, generalized critical points) and stratification theory. For nonsmooth analysis we refer to the comprehensive texts [5], [6], [28], [29], [32].

In what follows the vector space $\mathbb{R}^n$ is endowed with its canonical scalar product $\langle \cdot, \cdot \rangle$.

Nonsmooth analysis. Given an extended-real-valued function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ we denote its domain by $\operatorname{dom} f := \{x \in \mathbb{R}^n : f(x) < +\infty\}$, its graph by

$$\operatorname{Graph} f := \{(x, f(x)) \in \mathbb{R}^n \times \mathbb{R} : x \in \operatorname{dom} f\},$$

and its epigraph by

$$\operatorname{epi} f := \{(x, \beta) \in \mathbb{R}^n \times \mathbb{R} : f(x) \leq \beta\}.$$

In this work we shall deal with *lower semicontinuous* functions, that is, functions for which $\operatorname{epi} f$ is a closed subset of $\mathbb{R}^n \times \mathbb{R}$. In this setting, we say that $x^* \in \mathbb{R}^n$ is a *Fréchet subgradient* of f at $x \in \operatorname{dom} f$ provided that

$$\liminf_{y \to x, y \neq x} \frac{f(y) - f(x) - \langle x^*, y - x \rangle}{\|y - x\|} \geq 0. \tag{1}$$

The set of all Fréchet subgradients of f at x is called the *Fréchet subdifferential* of f at x and is denoted by $\hat{\partial} f(x)$. If $x \notin \operatorname{dom} f$, then we set $\hat{\partial} f(x) = \emptyset$.

Let us give a geometrical interpretation of the above definition: It is well known that the gradient of a C^1 function $f : \mathbb{R}^n \to \mathbb{R}$ at $x \in \mathbb{R}^n$ can be defined geometrically as the vector $\nabla f(x) \in \mathbb{R}^n$ such that $(\nabla f(x), -1)$ is normal to the tangent space $\mathrm{T}_{(x,f(x))}\mathrm{Graph}\, f$ of (the C^1 manifold) $\mathrm{Graph}\, f$ at $(x, f(x))$, that is,

$$(\nabla f(x), -1) \perp \mathrm{T}_{(x,f(x))}\mathrm{Graph}\, f.$$

A similar interpretation can be stated for Fréchet subgradients. Let us first define the *(Fréchet) normal cone* of a subset C of $\mathbb{R}^n$ at $x \in C$ by

$$\hat{N}_C(x) = \left\{ v \in \mathbb{R}^n : \limsup_{\substack{y \to x \\ y \in C \setminus \{x\}}} \left\langle v, \frac{y - x}{||x - y||} \right\rangle \leq 0 \right\}. \tag{2}$$

Then it can be proved (see [32, Theorem 8.9], for example) that for a nonsmooth function f we have

$$x^* \in \hat{\partial} f(x) \quad \text{if and only if} \quad (x^*, -1) \in \hat{N}_{\mathrm{epi}\, f}(x, f(x)). \tag{3}$$

The Fréchet subdifferential extends the notion of a derivative in the sense that if f is differentiable at x, then $\hat{\partial} f(x) = \{\nabla f(x)\}$. However, it is not completely satisfactory in optimization, since $\hat{\partial} f(x)$ might be empty-valued at points of particular interest (think of the example of the function $f(x) = -||x||$, at $x = 0$). Moreover, the Fréchet subdifferential is not a closed mapping, so it is unstable computationally. For this reason we also consider (see [28], [32], for example):

(i) the *limiting* subdifferential $\partial f(x)$ of f at $x \in \operatorname{dom} f$:

$$x^* \in \partial f(x) \iff \exists (x_n, x_n^*) \subset \mathrm{Graph}\, \hat{\partial} f \ : \ \begin{cases} \lim_{n\to\infty} x_n = x, \\ \lim_{n\to\infty} f(x_n) = f(x), \\ \lim_{n\to\infty} x_n^* = x^*, \end{cases} \tag{4}$$

where $\mathrm{Graph}\, \hat{\partial} f := \{(u, u^*) : u^* \in \hat{\partial} f(u)\}$;

(ii) the *singular limiting* subdifferential $\partial^\infty f(x)$ of f at $x \in \operatorname{dom} f$:

$$y^* \in \partial^\infty f(x) \iff \exists (y_n, y_n^*) \subset \mathrm{Graph}\, \hat{\partial} f,\ \exists t_n \searrow 0^+ : \begin{cases} \lim_{n\to\infty} y_n = x, \\ \lim_{n\to\infty} f(y_n) = f(x), \\ \lim_{n\to\infty} t_n y_n^* = y^*. \end{cases} \tag{5}$$

When $x \notin \operatorname{dom} f$ we set $\partial f(x) = \partial^\infty f(x) = \emptyset$.

The *Clarke subdifferential* $\partial^\circ f(x)$ of f at $x \in \operatorname{dom} f$ is the central notion of this work. It can be defined in several (equivalent) ways; see [5]. The definition below (see [16, Proposition 3.3], [17, Proposition 3.4], or [30, Theorem 8.11]) is the most convenient for our purposes. (For any subset S of $\mathbb{R}^n$ we denote by $\overline{\mathrm{co}}\, S$ the closed convex hull of S.)

DEFINITION 1 (Clarke subdifferential). *The Clarke subdifferential* $\partial^\circ f(x)$ *of* f *at* x *is the set*

$$\partial^\circ f(x) = \begin{cases} \overline{\text{co}}\,\{\partial f(x) + \partial^\infty f(x)\} & \text{if } x \in \operatorname{dom} f, \\ \emptyset & \text{if } x \notin \operatorname{dom} f\,. \end{cases} \tag{6}$$

Remark 1. The construction (6) does not look very natural at first sight. However, it can be shown that an analogous to (3) formula holds also for the Clarke subdifferential, if $\hat{N}_{\operatorname{epi} f}(x, f(x))$ is replaced by the Clarke normal cone, which is the closed convex hull of the limiting normal cone. The latter cone comes naturally from the Fréchet normal cone by closing its graph; and see [32, pp. 305 and 336] for details.

From the above definitions it follows directly that for all $x \in \mathbb{R}^n$, one has

$$\hat{\partial} f(x) \subset \partial f(x) \subset \partial^\circ f(x). \tag{7}$$

The elements of the limiting (respectively, Clarke) subdifferential are called limiting (respectively, Clarke) subgradients.

The notion of a Clarke critical point (respectively, critical value, asymptotic critical value) is defined as follows.

DEFINITION 2 (Clarke critical point). *We say that* $x \in \mathbb{R}^n$ *is a* Clarke critical point *of the function* f *if*

$$\partial^\circ f(x) \ni 0.$$

DEFINITION 3 ((asymptotic) Clarke critical value). (i) *We say that* $\alpha \in \mathbb{R}$ *is a* Clarke critical value *of* f *if the level set* $f^{-1}(\{\alpha\})$ *contains a Clarke critical point.*

(ii) *We say that* $\lambda \in \mathbb{R} \cup \{\pm\infty\}$ *is an* asymptotic Clarke critical value *of* f*, if there exists a sequence* $(x_n, x_n^*)_{n\geq 1} \subset \operatorname{Graph} \partial^\circ f$ *such that*

$$\begin{cases} f(x_n) \to \lambda \\ (1 + ||x_n||)\,||x_n^*|| \to 0. \end{cases}$$

Let us make some observations concerning the above definitions.

Remark 2. (i) Both limiting and Clarke subgradients are generalizations of the usual gradient of smooth functions: Indeed, if f is C^1 around x (or more generally, strictly differentiable at x [32, Definition 9.17]), then we have

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

It should be noted that if f is only Fréchet differentiable at x, then $\partial^\circ f(x) \supset \partial f(x) \supset \{\nabla f(x)\}$, where the inclusions might be strict.

(ii) The singular limiting subdifferential should not be thought as a set of subgradients. Roughly speaking it is designed to detect "horizontal normals" to the epigraph of f. For instance, for the (nonsmooth) function $f(x) = x^{\frac{1}{3}}$ $(x \in \mathbb{R})$ we have $\partial^\infty f(0) = \mathbb{R}_+$. Note that, since the domain of the Fréchet subdifferential is dense in $\operatorname{dom} f$, we always have $\partial^\infty f(x) \ni 0$ for all $x \in \operatorname{dom} f$ (see also [32, Corollary 8.10]); therefore, this latter relation cannot be regarded as a meaningful definition of a critical point.

(iii) To illustrate the definition of the Clarke critical point (Definition 1), let us consider the example of the function $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} x & \text{if } x \leq 0, \\ -\sqrt{x} & \text{if } x > 0. \end{cases}$$

Then $\hat{\partial} f(0) = \emptyset$ and $\partial f(0) = \{1\}$. However, since $\partial^\infty f(0) = \mathbb{R}_-$, it follows from (6) that $\partial^\circ f(0) = (-\infty, 1]$, so $x = 0$ is a Clarke critical point.

(iv) It follows from Definition 3 that every Clarke critical value $\alpha \in \mathbb{R}$ is also an asymptotic Clarke critical value (indeed, given $x_0 \in f^{-1}(\{\alpha\})$ with $0 \in \partial^\circ f(x_0)$, it is sufficient to take $x_n := x_0$ and $x_n^* = 0$). Note that in the case that f has a bounded domain $\operatorname{dom} f$, Definition 3(ii) can be simplified in the following way: The value $\lambda \in \mathbb{R} \cup \{\pm\infty\}$ is asymptotically critical if and only if there exists a sequence $(x_n, x_n^*)_{n\geq 1} \subset \operatorname{Graph} \partial^\circ f$ such that $f(x_n) \to \lambda$ and $x_n^* \to 0$.

Stratification results. By the term *stratification* we mean a locally finite partition of a given set into differentiable manifolds, which, roughly speaking, fit together in a regular manner. Let us give a formal definition of a C^p stratification of a set. For general facts about stratifications we quote [27]; more specific results concerning tame geometry can be found in [34], [11], [18], [10], [21].

Let X be a nonempty subset of $\mathbb{R}^n$ and p a positive integer. A C^p *stratification* $\mathcal{X} = (X_i)_{i\in I}$ of X is a locally finite partition of X into C^p submanifolds X_i of $\mathbb{R}^n$ such that for each $i \neq j$

$$\overline{X_i} \cap X_j \neq \emptyset \implies X_j \subset \overline{X_i} \setminus X_i.$$

The submanifolds X_i are called *strata* of $\mathcal{X}$. Furthermore, given a finite collection $\{A_1, \ldots, A_q\}$ of subsets of X, a stratification $\mathcal{X} = (X_i)_{i\in I}$ is said to be *compatible with the collection* $\{A_1, \ldots, A_q\}$ if each A_i is a locally finite union of strata X_j.

In this work we shall use a special type of stratification (called a *Whitney stratification*) for which the strata are such that their tangent spaces also "fit regularly." To give a precise meaning to this statement, let us first define the *distance* (or *gap*) of two vector subspaces V and W of $\mathbb{R}^n$ by the following standard formula:

$$D(V, W) = \max \left\{ \sup_{v\in V,\, ||v||=1} d(v, W), \sup_{w\in W,\, ||w||=1} d(w, V) \right\}.$$

Note that

$$\sup_{v\in V,\, ||v||=1} d(v, W) = 0 \Longleftrightarrow V \subset W.$$

Further we say that a sequence $\{V_k\}_{k\in\mathbb{N}}$ of subspaces of $\mathbb{R}^n$ *converges* to the subspace V of $\mathbb{R}^n$ (in short, $V = \lim_{k\to+\infty} V_k$) provided

$$\lim_{k\to+\infty} D(V_k, V) = 0.$$

Notice that in this case the subspaces V_k will eventually have the same dimension (say, d); thus, the above convergence is essentially equivalent to the convergence in the Grassmannian manifold G_d^n.

A C^p stratification $\mathcal{X} = (X_i)_{i\in I}$ of X has the *Whitney-*(a) *property,* if for each $x \in \overline{X_i} \cap X_j$ (with $i \neq j$) and for each sequence $\{x_k\} \subset X_i$ we have

$$\left.\begin{array}{ll} & \lim_{k\to\infty} x_k = x \\ \text{and} & \\ & \lim_{k\to\infty} T_{x_k} X_i = \mathcal{T}, \end{array}\right\} \implies T_x X_j \subset \mathcal{T},$$

where $T_x X_j$ (respectively, $T_{x_k} X_i$) denotes the tangent space of the manifold X_j at x (respectively, of X_i at x_k). In what follows we shall use the term *Whitney stratification* to refer to a C^1 stratification with the Whitney-(a) property.

3. Projection formulas for subgradients. In this section we make precise the links between the Clarke subgradients of a lower semicontinuous function whose graph admits a Whitney stratification and the gradients of f (with respect to the strata). As a corollary we obtain a nonsmooth extension of the Morse–Sard theorem for such functions (see Corollary 5).

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous function. We shall deal with *nonvertical* Whitney stratifications $\mathcal{S} = (S_i)_{i\in I}$ of the graph $\operatorname{Graph} f$ of f, that is, Whitney stratifications satisfying for all $i \in I$ and $u \in S_i$ the transversality condition

$$e_{n+1} \notin T_u S_i \qquad (\mathcal{H}),$$

where

$$e_{n+1} = (0, \dots, 0, 1) \in \mathbb{R}^{n+1}.$$

Remark 3. If f is locally Lipschitz continuous, then it is easy to check that any stratification of $\operatorname{Graph} f$ is nonvertical. This might also happen for other functions (think of the nonlocally Lipschitz function $f(x) = \sqrt{|x|}$: Every stratification of $\operatorname{Graph} f$ should contain the stratum $S_0 = \{(0,0)\}$). However, the example of the function $f(x) = x^{1/3}$ shows that this is not the case for any (continuous stratifiable) function f and any stratification of its graph (consider the trivial stratification consisting of the single stratum $S = \operatorname{Graph} f$ and take $u = (0,0)$).

Let us denote by $\Pi : \mathbb{R}^{n+1} \to \mathbb{R}^n$ the canonical projection on $\mathbb{R}^n$, that is,

$$\Pi(x_1, \dots, x_n, t) = (x_1, \dots, x_n).$$

For each $i \in I$ we set

$$X_i = \Pi(S_i) \qquad \text{and} \qquad f_i = f|_{X_i}. \tag{8}$$

Due to the assumption $(\mathcal{H})$ (nonverticality) one has that for all $i \in I$:

(i) X_i is a C^1 submanifold of $\mathbb{R}^n$, and

(ii) $f_i : X_i \to \mathbb{R}$ is a C^1 function.

If, in addition, the function f is continuous, then it can be easily seen that:

(iii) $\mathcal{X} = (X_i)_{i\in I}$ is a Whitney stratification of $\operatorname{dom} f = \Pi(\operatorname{Graph} f)$.

Notation. In what follows, for any $x \in \operatorname{dom} f$, we shall denote by X_x (respectively, S_x) the stratum of $\mathcal{X}$ (respectively, of $\mathcal{S}$) containing x (respectively, $(x, f(x))$). The manifolds X_i are here endowed with the metric induced by the canonical Euclidean scalar product of $\mathbb{R}^n$. Using the inherited Riemannian structure of each stratum X_i of $\mathcal{X}$ for any $x \in X_i$, we denote by $\nabla_R f(x)$ the gradient of f_i at x with respect to the stratum $X_i, \langle\cdot,\cdot\rangle$.

Proposition 4 (projection formula). *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous function, and assume that* $\operatorname{Graph} f$ *admits a nonvertical Whitney stratification $\mathcal{S} = (S_i)_{i\in I}$. Then for all $x \in \operatorname{dom} f$ we have*

$$\operatorname{Proj}_{T_x X_x} \partial f(x) \subset \{\nabla_{\mathrm{R}} f(x)\}, \qquad \operatorname{Proj}_{T_x X_x} \partial^\infty f(x) = \{0\}, \tag{9}$$

and

$$\operatorname{Proj}_{T_x X_x} \partial^\circ f(x) \subset \{\nabla_{\mathrm{R}} f(x)\}, \tag{10}$$

where $\operatorname{Proj}_{\mathcal{V}} : \mathbb{R}^n \to \mathcal{V}$ *denotes the orthogonal projection on the vector subspace $\mathcal{V}$ of $\mathbb{R}^n$.*

Proof. We shall use the above notation (and in particular the notation of (8)).

Let us first describe the links between the Fréchet subdifferential $\hat{\partial} f(x)$ and the gradient of $f\,|_{X_x}$ at a point $x \in \operatorname{dom} f$. For any $v \in T_x X_x$ and any continuously differentiable curve $c : (-\varepsilon, \varepsilon) \to X_x$ $(\varepsilon > 0)$ with $c(0) = x$ and $\dot{c}(0) = v$, the function

$$f \circ c \,(:= f_i \circ c) \,: (-\varepsilon, \varepsilon) \to \mathbb{R}$$

is continuously differentiable. In view of [32, Theorem 10.6, p. 427], we have

$$\left\{ \langle x^*,\, v\rangle :\, x^* \in \hat{\partial} f(x) \right\} \,\subset\, \left\{ \frac{d}{dt}\, f(c(t))\,|_{t=0} \right\}.$$

Since $\frac{d}{dt} f(c(t))|_{t=0} = \langle \nabla_R f(x), v\rangle$ it follows that

$$\operatorname{Proj}_{T_x X_x} \hat{\partial} f(x) \,\subset\, \{\nabla_R f(x)\}. \tag{11}$$

In a second stage we prove successively that

$$\operatorname{Proj}_{T_x X_x} \partial f(x) \,\subset\, \{\nabla_R f(x)\} \quad \text{and} \quad \operatorname{Proj}_{T_x X_x} \partial^\infty f(x) \,\subset\, \{0\}. \tag{12}$$

To this end, take $p \in \partial f(x)$, and let $\{x_k\} \,\subset\, \operatorname{dom} \hat{\partial} f$, $x_k^* \in \hat{\partial} f(x_k)$ be such that $(x_k,\, f(x_k)) \to (x,\, f(x))$ and $x_k^* \to p$. Due to the local finiteness property of $\mathcal{S}$, we may suppose that the sequence $\{u_k := (x_k, f(x_k))\}$ lies entirely in some stratum S_i of dimension d.

If $S_i = S_x$, then by (11) we deduce that $\operatorname{Proj}_{T_x X_x}(x_k^*) = \nabla_R f(x_k)$; thus, using the continuity of the projection and the fact that $f\,|_{X_x}$ is C^1 (so $\nabla_R f(x_k) \to \nabla_R f(x)$), we obtain $\operatorname{Proj}_{T_x X_x}(p) = \nabla_R f(x)$.

If $S_i \neq S_x$, then from the convergence $(x_k,\, f(x_k)) \to (x,\, f(x))$ we deduce that $\overline{S_i} \cap S_x \neq \emptyset$ (thus $d = \dim S_i > \dim S_x$). Using the compactness of the Grassmannian manifold G_d^n, we may assume that the sequence $\{T_{u_k} S_i\}$ converges to some vector space $\mathcal{T}$ of dimension d. Then the Whitney-(a) property yields that $\mathcal{T} \supset T_{(x,f(x))} S_x$. Recalling (3), for each $k \geq 1$ we have that the vector $(x_k^*, -1)$ is Fréchet normal to the epigraph $\operatorname{epi} f$ of f at u_k; hence, it is also normal (in the classical sense) to the tangent space $T_{u_k} S_i$. By a standard continuity argument the vector

$$(p, -1) = \lim_{k \to \infty} (x_k^*, -1)$$

must be normal to $\mathcal{T}$ and a fortiori to $T_{(x,f(x))} S_x$. By projecting $(p, -1)$ orthogonally on $T_x X_x + \mathbb{R}\, e_{n+1} \supset T_{(x,f(x))} S_x$, we notice that $(\operatorname{Proj}_{T_x X_x}(p), -1)$ is still normal to $T_{(x,f(x))} S_x$. We conclude that

$$\operatorname{Proj}_{T_x X_x}(p) = \nabla_R f(x); \tag{13}$$

thus, the first part of (12) follows.

Let now any $q \in \partial^\infty f(x)$. By definition there exist $\{y_k\} \,\subset\, \operatorname{dom} \hat{\partial} f$, $y_k^* \in \hat{\partial} f(y_k)$, and a positive sequence $t_k \searrow 0^+$ such that $(y_k,\, f(y_k)) \to (y,\, f(y))$ and $t_k y_k^* \to q$. As above we may assume that the sequence $\{y_k\}$ belongs to some stratum S_i and that the tangent spaces $T_{u_k} S_i = T_{(x_k, f(x_k))} S_i$ converge to some $\mathcal{T}$. Since $t_k(y_k^*, -1)$ is normal to $T_{u_k} S_i$ we can similarly deduce that $(\operatorname{Proj}_{T_x X_x}(q),\, 0)$ is normal to $T_{(x,f(x))} S_x$. Since $\operatorname{Proj}_{\mathbb{R}^n \times \{0\}} T_{(x,f(x))} S_x = T_x X_x$ this implies that $\partial^\infty f(x) \subset (T_x X_x)^\perp$, and the second part of (12) is proved. It now follows from (12) and Remark 2(ii) that (9) holds.

In order to conclude let us recall (Definition 1) that $\partial^\circ f(x) = \overline{\text{co}}\,(\partial f(x)+\partial^\infty f(x))$. In view of (12) any element of co $(\partial f(x) + \partial^\infty f(x))$ admits $\nabla_R f(x)$ as a projection onto $T_x X_x$. By taking the closure of the previous set we obtain (10). □

Remark 4. The inclusion in (10) may be strict (think of the function $f(x) = -||x||^{1/2}$ at $x = 0$, where $\partial^\circ f(0) = \emptyset$). Of course, whenever $\partial^\circ f(x)$ is nonempty (for example, if f is locally Lipschitz), under the assumptions of Proposition 4 we have

$$\text{Proj}_{T_x X_x} \partial^\circ f(x) = \{\nabla_R f(x)\}.$$

COROLLARY 5. *Assume that f is lower semicontinuous and admits a nonvertical C^p-Whitney stratification. Then:*

(i) *for all $x \in \text{dom}\,\partial^\circ f$ we have*

$$||\nabla_{\text{R}} f(x)|| \;\le\; ||x^*|| \qquad \textit{for all } x^* \in \partial^\circ f(x). \tag{14}$$

(ii) (Morse–Sard theorem) *If $p \ge n$, then the set of Clarke critical values of f has Lebesgue measure zero.*

Proof. Assertion (i) is a direct consequence of (10) of Proposition 4. To prove (ii), set $C := [\partial^\circ f]^{-1}(\{0\}) = \{x \in \mathbb{R}^n : \partial^\circ f(x) \ni 0\}$. Since the set of strata is at most countable, the restrictions of f to each of those yield a countable family $\{f_n\}_{n\in\mathbb{N}}$ of C^p functions. In view of (14), we have that $C \subset \cup_{n\in\mathbb{N}}(\nabla f_n)^{-1}(0)$. The result follows by applying to each C^p-function f_n the classical Morse–Sard theorem [33]. □

As we see in the next section, several important classes of lower semicontinuous functions satisfy the assumptions (thus also the conclusions) of Proposition 4 and of Corollary 5.

4. Kurdyka–Łojasiewicz inequalities for o-minimal functions. Let us recall briefly a few definitions concerning o-minimal structures (see, for instance, Coste [7], van den Dries and Miller [11], Ta Lê Loi [35], and references therein).

DEFINITION 6 (o-minimal structure). *An o-minimal structure on $(\mathbb{R}, +, .)$ is a sequence of Boolean algebras $\mathcal{O}_n$ of "definable" subsets of $\mathbb{R}^n$ such that for each $n \in \mathbb{N}$:*

(i) if A belongs to $\mathcal{O}_n$, then $A \times \mathbb{R}$ and $\mathbb{R} \times A$ belong to $\mathcal{O}_{n+1}$;

(ii) if $\Pi : \mathbb{R}^{n+1} \to \mathbb{R}^n$ is the canonical projection onto $\mathbb{R}^n$, then for any A in $\mathcal{O}_{n+1}$ the set $\Pi(A)$ belongs to $\mathcal{O}_n$;

(iii) $\mathcal{O}_n$ contains the family of algebraic subsets of $\mathbb{R}^n$, that is, every set of the form

$$\{x \in \mathbb{R}^n : p(x) = 0\},$$

where $p : \mathbb{R}^n \to \mathbb{R}$ is a polynomial function;

(iv) the elements of $\mathcal{O}_1$ are exactly the finite unions of intervals and points.

DEFINITION 7 (definable function). *Given an o-minimal structure $\mathcal{O}$ (over $(\mathbb{R}, +, .)$), a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is said to be* definable *in $\mathcal{O}$ if its graph belongs to $\mathcal{O}_{n+1}$.*

Remark 5 (examples). At first sight, o-minimal structures might appear artificial in optimization. The following fundamental properties (see [11] for the details) might convince the reader that this is not the case.

(i) (Tarski–Seidenberg) The collection of *semialgebraic sets* is an o-minimal structure. Recall that semialgebraic sets are Boolean combinations of sets of the form

$$\{x \in \mathbb{R}^n : p(x) = 0,\, q_1(x) < 0, \ldots,\, q_m(x) < 0\},$$

where p and q_i's are polynomial functions on $\mathbb{R}^n$.

(ii) (Gabrielov) There exists an o-minimal structure that contains the sets of the form

$$\{(x,t) \in [-1,1]^n \times \mathbb{R} : f(x) = t\},$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is real analytic around $[-1,1]^n$.

(iii) (Wilkie) There exists an o-minimal structure that contains simultaneously the graph of the exponential function $\mathbb{R} \ni x \mapsto \exp x$ and all semialgebraic sets (respectively, all sets of the structure defined in (ii)).

We insist on the fact that these results are crucial foundation blocks on which o-minimal geometry rests.

Let us finally recall the following elementary but important result: The composition of mappings that are definable in some o-minimal structure remains in the same structure [11, section 2.1]. This is also true for the sum, the inf-convolution, and several other classical operations of analysis involving a finite number of definable objects. Another prominent fact about definable sets is that they admit, for each $k \geq 1$, a C^k–Whitney stratification with finitely many strata (see, for instance, [11, Result 4.8, p. 510]). This remarkable stability, combined with new techniques of finite-dimensional optimization, offers a large field of investigation. Several works have already been developed in this spirit; see, for instance, [1], [3], [12].

Given any o-minimal structure $\mathcal{O}$ and any lower semicontinuous definable function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, the assumptions of Proposition 4 are satisfied. More precisely, we have the following result.

LEMMA 8. *Let $\mathcal{O}$ be an o-minimal structure, $\mathcal{B} := \{B_1, \ldots, B_q\}$ be a collection of definable subsets of $\mathbb{R}^n$, and $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a definable lower semicontinuous function. Then for any $p \geq 1$, there exists a nonvertical definable C^p-Whitney stratification $\{S_1, \ldots, S_\ell\}$ of* Graph f *yielding (by projecting each stratum $S_i \subset \mathbb{R}^{n+1}$ onto $\mathbb{R}^n$) a C^p-Whitney stratification $\{X_1, \ldots, X_\ell\}$ of* dom f *compatible with $\mathcal{B}$.*

Proof. By transforming, using diffeomorphisms preserving verticality, $\mathbb{R}^n$ to $D := \{x \in \mathbb{R}^n : ||x|| < 1\}$ and $\mathbb{R}$ to $(-1,1)$, we may assume without loss of generality that f is defined in $D := \{x \in \mathbb{R}^n : ||x|| < 1\}$ with values in $(-1,1)$. Set $X =$ Graph f and $A_i = B_i \times (-1,1)$ for $i \in \{1, \ldots, q\}$, and let $\pi : X \to D$ denote the restriction to Graph f of the canonical projection of $D \times (-1,1)$ to D. The lemma follows from the canonical stratification of the mapping π according to [34, II.1.17]. ∎

COROLLARY 9 (Morse–Sard theorem for definable functions). *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous definable function and $p \geq 1$. Then there exists a finite definable C^p-Whitney stratification $\mathcal{X} = (X_i)_{i \in I}$ of* dom f *such that for all $x \in$ dom f*

$$\text{Proj}_{T_x X_x} \partial^\circ f(x) \subset \{\nabla_{\mathrm{R}} f(x)\}. \tag{15}$$

As a consequence:

(i) *for all $x \in$ dom $\partial^\circ f$ and $x^* \in \partial^\circ f(x)$, we have $||\nabla_{\mathrm{R}} f(x)|| \leq ||x^*||$;*

(ii) *the set of Clarke critical values of f is finite;*

(iii) *the set of asymptotic Clarke critical values of f is finite.*

Proof. Assertion (i) is a direct consequence of (15). This projection formula follows directly by combining Lemma 8 with Proposition 4. To prove (iii), let f_i be the restriction of f to the stratum X_i. Then assertion (i), together with the fact that the number of strata is finite, implies that the set of the asymptotic Clarke critical values of f is the union (over the finite set I) of the asymptotic critical values of

each (definable C^1) function f_i. Thus the result follows from [8, Remarque 3.1.5]. Assertion (ii) follows directly from (iii) (cf. Remark 2(iii)). ∎

Remark 6. The fact that the set of the asymptotic critical values of a definable differentiable function f is finite has been established in [8, Théorème 3.1.4] (see also [20, Theorem 3.1] for the case that the domain of f is bounded). In [22, Proposition 2] a more general result (concerning functions taking values in $\mathbb{R}^k$) has been established in the semialgebraic case.

We shall now give another application of Proposition 4, namely, a nonsmooth version of the classical Kurdyka–Łojasiewicz inequality ([20, Theorem 1]). Before we proceed, we shall improve the latter in a way that allows us to deal directly with unbounded domains. To this end, we shall need the following proposition.

PROPOSITION 10 (uniform boundedness). *Let $I = [a, +\infty)$ for some $a \in \mathbb{R}$, and let $\mathcal{V}$ be a definable neighborhood of $\{0\} \times I$ in $\mathbb{R}_+ \times I$ and $\phi : \mathcal{V} \to \mathbb{R}_+$ a definable function, continuous throughout $\{0\} \times I$, satisfying $\phi(0, s) = 0$ for all $s \in I$. Then there exist $\varepsilon_0 > 0$ and continuous definable functions $\chi : I \to (0, \varepsilon_0)$ and $\psi : (0, \varepsilon_0) \to [0, +\infty)$ such that ψ is C^1 on $(0, \varepsilon_0)$, $\psi(0) = 0$, and*

$$\psi(t) \geq \phi(t, s) \quad \text{for all } s \in I, \qquad t \in (0, \chi(s)).$$

Proof. We can clearly assume that $a = 0$. Since $\mathcal{V}$ is a definable neighborhood of $\{0\} \times I$, we may assume there exists a continuous definable function $g : I \to (0, +\infty)$ such that $\{(t, s) \in \mathbb{R}_+ \times I : t \leq g(s)\} \subset \mathcal{V}$. Set

$$\delta(s) := \sup \left\{ \delta \in (0, g(s)) : \quad \phi(t, s) \leq \frac{1}{s+1} \quad \forall t \in [0, \delta) \right\}, \tag{16}$$

and note that $\delta(s)$, being definable, has a finite number of points of discontinuity. Since ϕ is continuous on $\{0\} \times I$ and $\phi(0, s) = 0$ for all $s \in I$, we infer that $\liminf_{s \to \bar{s}} \delta(s) > 0$ for all $\bar{s} \in I$. We deduce that there exists a continuous decreasing and definable function $\chi : I \to (0, +\infty)$ satisfying $\chi(s) \leq \delta(s)$ for all $s \in I$. Set $\varepsilon_0 = \sup \chi(I) = \chi(0) > 0$, and consider the definable function

$$\psi(t) = \max_{s \in [0, \chi^{-1}(t)]} \phi(t, s) \quad \text{for all } t \in [0, \varepsilon_0).$$

By the monotonicity lemma [7, Theorem 2.1] we conclude that ψ is C^1 on $(0, \beta)$ for some $\beta \leq \varepsilon_0$. Truncating χ if necessary (by defining $\tilde{\chi}(s) := \min\{\beta, \chi(s)\}$), we see that there is no loss of generality to assume $\beta = \varepsilon_0$. Note that $\psi(0) = 0$. Let us show that ψ is also continuous at $t = 0$. Let us assume, towards a contradiction, that there exists a sequence $t_n \searrow 0^+$ satisfying $\psi(t_n) > c > 0$. Then for every $n \in \mathbb{N}$ there exists $s_n \in [0, \chi^{-1}(t_n)]$ such that $\phi(t_n, s_n) > c > 0$. If $\{s_n\} \to +\infty$, then since $\delta(s_n) \geq \chi(s_n)$ we would deduce from (16) that $(s_n + 1)^{-1} \geq \phi(t_n, s_n) > c$, which is impossible for large values of n. Thus $\{s_n\}$ is bounded and has a convergent subsequence to some $s \in I$. Using the continuity of ϕ at $(0, s)$ and the fact that $\phi(0, s) = 0$, the contradiction follows. One can easily check that the definable functions ψ and χ satisfy the conclusion of the proposition. ∎

We now provide the following extension of the Kurdyka–Łojasiewicz inequality ([20, Theorem 1]) for unbounded sets in the smooth case.

THEOREM 11 (Kurdyka–Łojasiewicz inequality). *Let U be a nonempty definable submanifold of $\mathbb{R}^n$ (not necessarily bounded) and $f : U \to \mathbb{R}_+$ be a definable differentiable function. Then there exist a continuous definable function $\psi : [0, \varepsilon_0) \to \mathbb{R}_+$ satisfying $\psi(0) = 0$ and being C^1 on $(0, \varepsilon_0)$ and a continuous definable function $\chi : \mathbb{R}_+ \to (0, \varepsilon_0)$ such that*

$$||\nabla(\psi \circ f)(x)|| \geq 1 \quad \text{for all } 0 < f(x) \leq \chi(||x||). \tag{17}$$

Proof. With no loss of generality we can assume that f is not identically equal to 0 on U.

For each $(t, s) \in (0, +\infty) \times \mathbb{R}_+$ we set

$$F(t,s) := f^{-1}(t) \cap B(0,s) \subset U \qquad \text{and} \qquad m_f(t,s) = \inf\ \{||\nabla f(x)|| : x \in F(t,s)\}. \tag{18}$$

Note that $m_f(t, s) \equiv +\infty$ whenever $F(t, s)$ is empty. If $f^{-1}(0) = \emptyset$, then for every $s \geq 0$ there exists $\delta > 0$ such that for all $t \in (0, \delta)$ we have $F(t, s) = \emptyset$. Thus, the definable function

$$s \mapsto \delta(s) := \sup\{\delta > 0 : F(t,s) = \emptyset\ \ \forall t \in (0, \delta]\} < +\infty$$

is positive (cf. continuity of f), decreasing (since $F(t, s_1) \subset F(t, s_2)$ for $s_1 \leq s_2$), and continuous on $(\bar{s}, +\infty)$ for some $\bar{s} > 0$ (cf. monotonicity lemma [7, Theorem 2.1]). In this case (17) follows trivially by considering the continuous function

$$\chi(s) = \begin{cases} \delta(s)/2 & \text{if } s \geq \bar{s}, \\ \delta(\bar{s})/2 & \text{if } s \leq \bar{s}, \end{cases}$$

and any continuous definable function ψ.

Thus there is no loss of generality to assume that there exists $s_0 \geq 0$ and a decreasing continuous definable function $\rho : [s_0, +\infty) \to (0, +\infty)$ such that $F(t, s) \neq \emptyset$ for all $t \in [0, \rho(s)]$ and all $s \geq s_0$. It follows that for all $s \geq s_0$ and $t \in [0, \rho(s)]$ we have $m_f(t, s) \in \mathbb{R}_+$ and (since $\arg\min f = \{0\}$) $m_f(0, s) = 0$. Using an argument of Kurdyka ([20, Claim, p. 777]) we deduce that the function $t \mapsto m_f(t, s)$ is not identically 0 near the origin, and we set for all $s \geq s_0$

$$g(s) = \sup\ \{\, t_0 \in (0, \rho(s)) :\ m_f(t,s) > 0\ \forall t \in (0, t_0]\,\} \in (0, +\infty).$$

Then g is decreasing, positive, definable, and thus continuous on $[s_1, +\infty)$ for some $s_1 \geq s_0$. Set $D = \{(t, s) \in \mathbb{R}_+ \times [s_1, +\infty) : t \leq g(s)\}$, and consider the following definable point-to-set mapping $M : D \rightrightarrows U \subset \mathbb{R}^n$, with

$$M(t,s) := \{x \in F(t,s) : ||\nabla f(x)|| \leq\ 2\ m_f(t,s)\}.$$

Using the definable selection lemma (cf. [7, Theorem 3.1]), we obtain a definable mapping $\gamma : D \to \mathbb{R}^n$ such that $\gamma(t, s) \in M(t, s)$ for all $(t, s) \in D$. Note that for each s fixed, the function $(0, g(s)) \ni t \mapsto \gamma(t, s)$ is absolutely continuous and $\frac{\partial}{\partial t}\gamma_i(\cdot, s)$ changes sign only a finite number of times on D for all $i \in \{1, \ldots, n\}$. We set

$$\phi(t,s) = \int_0^t \max_{i \in \{1,\ldots,n\}} \left| \frac{\partial}{\partial t}\gamma_i(\tau, s) \right| \, d\tau$$

for all $(t, s) \in D$. Applying the monotonicity lemma we obtain the integrability of the function

$$\tau \mapsto \max_{i=1,\ldots,n} \left| \frac{\partial}{\partial t}\gamma_i(\tau, s) \right|.$$

Using routine arguments it is easily seen that ϕ is actually definable on D. Moreover, $\phi(t,s) > 0$ whenever $t > 0$ (else the curve $\gamma(\cdot, s)$ would be stationary, which is not possible since $f(\gamma(t,s)) = t$). Note also that $\phi(0,s) = 0$ and $\lim_{t \searrow 0^+} \phi(t,s) = 0$. Considering a stratification of ϕ we deduce that there exists $a \geq s_1$ and a definable neighborhood $\mathcal{V}$ of $\{0\} \times [a, +\infty)$ in D where ϕ is (jointly) continuous. Applying Proposition 10, we obtain $\varepsilon_0 > 0$, a continuous definable function $\chi : [a, +\infty) \to (0, \varepsilon_0)$, and a continuous definable function $\psi : [0, \varepsilon_0) \to \mathbb{R}$, with $\psi(0) = 0$, such that ψ is C^1 on $(0, \varepsilon_0)$ and $\psi(t) \geq \phi(t,s)$ for all $t \in [0, \chi(s)]$.

Fix $s \geq a$. Since $\psi(t) \geq \phi(t,s)$ for $t \in [0, \chi(s)]$ and $\psi(0) = \phi(0,s)$, it follows (see [2, Lemma 1(i)], for example) that for all $t > 0$ sufficiently small

$$\psi'(t) \geq \frac{\partial}{\partial t}\phi(t,s) > 0. \tag{19}$$

For each $s \in [a, +\infty)$ let us define $\varepsilon(s)$ to be the supremum of all $\varepsilon \in (0, \varepsilon_0)$ such that (19) holds true in the interval $(0, \varepsilon)$. It follows that $s \mapsto \varepsilon(s)$ is a positive definable function and is thus continuous on $[b, +\infty)$ for some $b \geq a$. Let us define

$$\tilde{\chi}(s) = \begin{cases} \min\{\chi(s), \varepsilon(s)\} & \text{if } s \geq b, \\ \min\{\chi(b), \varepsilon(b)\} & \text{if } s \in [0, b]. \end{cases}$$

We shall now show that (17) holds for $\tilde{\psi} = (\frac{1}{2\sqrt{n}})\psi$ and for $\tilde{\chi} : \mathbb{R}_+ \to (0, \varepsilon_0)$. Indeed, let $x \in U$ be such that $0 < f(x) \leq \tilde{\chi}(||x||)$ (hence $\nabla f(x) \neq 0$). Set $t = f(x)$ and $s = \max\{||x||, b\}$. Using the definition of γ we obtain

$$||\nabla(\psi \circ f)(x)|| = \psi'(t)||\nabla f(x)|| \geq \frac{1}{2}\psi'(t)||\nabla f(\gamma(t,s))||. \tag{20}$$

On the other hand, since $f(\gamma(t,s)) = t$, we have

$$\frac{d}{dt} f(\gamma(t,s)) = \left\langle \frac{\partial}{\partial t}\gamma(t,s), \nabla f(\gamma(t,s)) \right\rangle = 1$$

for all $(t,s) \in D$; hence

$$\sqrt{n} \max_{i=1,\ldots,n} \left| \frac{\partial}{\partial t}\gamma_i(\cdot, s) \right| \; ||\nabla f(\gamma(t,s))|| \geq \left\| \frac{\partial}{\partial t}\gamma(t,s) \right\| \; ||\nabla f(\gamma(t,s))|| \geq 1,$$

and thus

$$||\nabla f(\gamma(t,s))|| \geq \left[\sqrt{n} \max_{i=1,\ldots,n} \left| \frac{\partial}{\partial t}\gamma_i(\cdot, s) \right| \right]^{-1} = \left[\sqrt{n} \frac{\partial}{\partial t}\phi(t,s) \right]^{-1}. \tag{21}$$

Since $f(x) \leq \tilde{\chi}(||x||) \leq \varepsilon(s)$, by combining (19), (20), and (21) we finally obtain that

$$||\nabla(\psi \circ f)(x)|| \geq \frac{1}{2\sqrt{n}}\psi'(t) \left[\frac{\partial}{\partial t}\phi(t,s) \right]^{-1} \geq \frac{1}{2\sqrt{n}};$$

that is, (17) holds for $\tilde{\psi} = (\frac{1}{2\sqrt{n}})^{-1}\psi$. □

Remark 7. If in the statement of Theorem 11 the definable set U is not open, then ∇f is understood as the Riemannian gradient of f on U.

We easily obtain the following corollaries.

COROLLARY 12. *Let $f : U \to \mathbb{R}$ be a definable differentiable function, where U is a definable submanifold of $\mathbb{R}^n$ (not necessarily bounded). Then there exist a continuous definable function $\psi : [0, \varepsilon_0) \to \mathbb{R}_+$ which is C^1 on $(0, \varepsilon_0)$, with $\psi(0) = 0$, and a relatively open neighborhood V of $f^{-1}(0)$ in U such that*

$$||\nabla(\psi \circ |f|)(x)|| \geq 1$$

for all x in $V \setminus f^{-1}(0)$.

Proof. Let us first assume that f is nonnegative. The result holds trivially if $f^{-1}(0) = \emptyset$, so let us assume $f^{-1}(0) \neq \emptyset$. Take ψ and χ as in Theorem 11, and let $x \in f^{-1}(0)$. It suffices to show that the inequality holds on a ball around x. Take $r \in (0, \varepsilon_0)$ such that $\chi(||x||) > r$. Since χ and f are continuous, there exists $\delta > 0$ such that $y \in B(x, \delta) \cap U$ implies $\chi(||y||) > r > f(y)$. Applying Theorem 11, we conclude that for all $y \in B(x, \delta) \cap U$ inequality (17) holds. When f takes its values in $\mathbb{R}$ (not necessarily in $\mathbb{R}_+$), the conclusion follows easily by considering the submanifolds $\{x \in U : f(x) > 0\}$, $\{x \in U : f(x) < 0\}$ and by applying the monotonicity Lemma. ∎

COROLLARY 13. *Let $f : U \to \mathbb{R}$ be a definable differentiable function, where U is a definable submanifold of $\mathbb{R}^n$ (not necessarily bounded). Let us denote by $C_1, \ldots, C_m$ the connected components of $(\nabla f)^{-1}(\{0\})$ and by $c_1, \ldots, c_m$ the corresponding critical values. Then there exist a continuous definable function $\psi : [0, \varepsilon_0) \to \mathbb{R}_+$ which is C^1 on $(0, \varepsilon_0)$, with $\psi(0) = 0$, and relatively open neighborhoods V_i of C_i in U for each $i \in \{1, \ldots, m\}$ such that for all $x \in V_i \setminus C_i$ we have*

$$||\nabla[\psi \circ |f - c_i|](x)|| \geq 1.$$

Proof. Note that $(\nabla f)^{-1}(\{0\}) \subset \cup_{i=1}^m f^{-1}(c_i)$. For each $i \in \{1, \ldots, m\}$ we apply Corollary 12 to the function $f_i := f - c_i$ on U to obtain a relatively open neighborhood V_i of C_i and $\psi_i : [0, \varepsilon_i) \to \mathbb{R}_+$ such that for all $x \in V_i \setminus f^{-1}(c_i)$

$$||\nabla[\psi_i \circ |f - c_i|](x)|| \geq 1.$$

Set $\varepsilon_0 = \min\{\varepsilon_i : i \in \{1, \ldots, m\}\}$. Since ψ_i are definable functions, shrinking ε_0 if necessary, we may assume (cf. the monotonicity lemma) that $\psi'_{i_0}(t) \geq \psi'_i(t)$ for all $t \in (0, \varepsilon_0)$ and all $i \in \{1, \ldots, m\}$. The conclusion follows by setting $\psi := \psi_{i_0}$ on $[0, \varepsilon_0)$. ∎

We shall now use Corollary 9 to extend Theorem 11 to a nonsmooth setting.

THEOREM 14 (nonsmooth Kurdyka–Łojasiewicz inequality). *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous definable function. There exist $\rho > 0$, a strictly increasing continuous definable function $\psi : [0, \rho) \to (0, +\infty)$ which is C^1 on $(0, \rho)$, with $\psi(0) = 0$, and a continuous definable function $\chi : \mathbb{R}_+ \to (0, \rho)$ such that*

$$||x^*|| \geq \frac{1}{\psi'(|f(x)|)}, \tag{22}$$

whenever $0 < |f(x)| \leq \chi(||x||)$ and $x^* \in \partial^\circ f(x)$.

Proof. Set $U_1 = \{x \in \operatorname{dom} f : f(x) > 0\}$ and $U_2 = \{x \in \operatorname{dom} f : f(x) < 0\}$, and let $X_1, \ldots, X_l$ be a finite definable stratification of $\operatorname{dom} f$ compatible with the (definable) sets U_1 and U_2 such that the definable sets $S_i = \{(x, f(x)) : x \in X_i\}$ are the strata of a nonvertical definable C^p-Whitney stratification of $\operatorname{Graph} f$ (cf. Lemma 8). For each $i \in \{1, \ldots, l\}$ such that $X_i \subset U_1$ we consider the positive

C^1 function $f_i := f|_{X_i}$ on the definable manifold X_i (thus for $x \in X_i$ we have $\nabla f_i(x) = \nabla_R f(x)$ and $f_i(x) = f(x)$), and we apply Theorem 11 to obtain $\varepsilon_i > 0$, a continuous definable function $\chi_i : \mathbb{R}_+ \to (0, \varepsilon_i)$, and a strictly increasing definable C^1-function $\psi_i : (0, \varepsilon_i) \to (0, +\infty)$ such that for all $x \in f^{-1}(0, \chi_i(||x||))$ we have $||\nabla_R f(x)|| \geq [\psi_i'(f(x))]^{-1}$. Similarly, for each $j \in \{1, \ldots, l\}$ such that $X_j \subset U_2$ we consider the positive C^1-function $f_j := -f|_{X_i}$ (note that for $x \in X_j$ we have $\nabla f_j(x) = -\nabla_R f(x)$ and $f_j(x) = -f(x)$) to obtain as before a definable function $\chi_j : \mathbb{R}_+ \to (0, \varepsilon_j)$ and a strictly increasing definable C^1-function $\psi_j : (0, \varepsilon_j) \to (0, +\infty)$ such that for all $x \in f^{-1}(0, \chi_i(||x||))$ we have $||\nabla_R f(x)|| \geq [\psi_j'(-f(x))]^{-1}$. Thus for all $i \in \{1, \ldots, l\}$ there exist a definable function $\chi_i : \mathbb{R}_+ \to (0, \varepsilon_i)$ and a strictly increasing definable C^1-function $\psi_i : (0, \varepsilon_i) \to \mathbb{R}$ such that

$$||\nabla_R f(x)|| \geq \frac{1}{\psi_i'(|f(x)|)} \qquad \text{for all } x \in f^{-1}(0, \chi_i(||x||)).$$

Set $\chi = \min \chi_i$, $\rho = \min \varepsilon_i$, and let $i_1, i_2 \in \{1, \ldots, l\}$. By the monotonicity theorem for definable functions of one variable (see [20, Lemma 2], for example), the definable function

$$(0, \rho) \ni r \quad \mapsto \quad 1/\psi_{i_1}'(r) - 1/\psi_{i_2}'(r)$$

has a constant sign in a neighborhood of 0. Repeating the argument for all couples i_1, i_2 and shrinking ρ if necessary, we obtain the existence of a strictly increasing, positive, definable function $\psi = \psi_{i_0}$ on $(0, \rho)$ of class C^1 that satisfies $1/\psi' \leq 1/\psi_i'$ on $(0, \rho)$ for all $i \in \{1, \ldots, l\}$. Evoking Corollary 9(i), we obtain

$$||x^*|| \geq ||\nabla_R f(x)|| \geq \frac{1}{\psi'(|f(x)|)},$$

whenever $x \in |f|^{-1}(0, \chi(||x||))$ and $x^* \in \partial^\circ f(x)$. Since ψ is definable and bounded from below, it can be extended continuously to $[0, \rho)$. By eventually adding a constant, we can also assume $\psi(0) = 0$. □

In a similar way to Corollary 13 we obtain the following result.

COROLLARY 15. *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous definable function. Let us denote by $C_1, \ldots, C_m$ the connected components of $(\partial^\circ f)^{-1}(\{0\})$ and by $c_1, \ldots, c_m$ the corresponding critical values (cf. Corollary* 9(ii)*). Then there exist a continuous definable function $\psi : [0, \varepsilon_0) \to \mathbb{R}_+$ which is C^1 on $(0, \varepsilon_0)$, with $\psi(0) = 0$, and relatively open neighborhoods V_i of C_i in* dom f *for each $i \in \{1, \ldots, m\}$ such that for all $x \in V_i$ we have*

$$||x^*|| \geq \frac{1}{\psi'(|f(x) - c_i|)}, \tag{23}$$

whenever $0 < |f(x) - c_i| \leq \chi(||x||)$ and $x^* \in \partial^\circ f(x)$.

The assumption that the function f is definable is important for the validity of (22). It implies in particular that the connected components of the set of the Clarke critical points of f lie in the same level set of f (cf. Corollary 9(ii)). Let us present some examples of C^1-functions for which (22) is not true.

Example 1. (i) Consider the function $f : \mathbb{R} \to \mathbb{R}$, with

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

Then the set $S = \{x \in \mathbb{R} : f'(x) = 0\}$ meets infinitely many level sets around 0. Consequently, (22) is not fulfilled since the critical value 0 is not isolated. Note also that f provides an example of a nondefinable function whose graph admits a Whitney stratification (in particular f satisfies the conclusion of Proposition 4).

(ii) A nontrivial example is proposed in [31, p. 14], where a C^∞ "Mexican-hat" function has been defined. An example of a similar nature has been given in [1] and will be described below: Let f be defined in polar coordinate on $\mathbb{R}^2$ by

$$f(r,\theta) = \begin{cases} \exp(-\frac{1}{1-r^2})\,[1 - \frac{4r^4}{4r^4+(1-r^2)^4}\,\sin(\theta - \frac{1}{1-r^2})] & \text{if } r \le 1, \\ \\ 0 & \text{if } r > 1. \end{cases}$$

The function f does not satisfy the Kurdyka–Łojasiewicz inequality for the critical value 0; i.e., one cannot find a strictly increasing C^1-function $\psi : (0,\rho) \to (0,+\infty)$, with $\rho > 0$, such that

$$||\nabla(\psi \circ f)(x)|| \ge 1$$

for small positive values of $f(x)$. To see this, let us notice that the proof of [20, Theorem 2] shows that for *any* C^1-function f (not necessarily definable) that satisfies the Kurdyka–Łojasiewicz inequality, the bounded trajectories of the gradient system

$$\dot{x}(t) + \nabla f(x(t)) = 0$$

have a bounded length. However, in the present example, taking as the initial condition $r_0 \in (0,1)$ and θ_0 such that $\theta_0(1-r_0)^2 = 1$, the gradient trajectory $\dot{x}(t) = -\nabla f(x(t))$ must comply with

$$\theta(t) = \frac{1}{1-r(t)^2},$$

where $r(t) \nearrow 1^-$ as $t \to +\infty$ (see [1] for details). The total length of the above curve is obviously infinite, which shows that the Kurdyka–Łojasiewicz inequality (for the critical value 0) does not hold.

Let us finally give an easy consequence of Theorem 14 for the case of subanalytic functions [25].

COROLLARY 16 (subgradient inequality). *Assume that $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a lower semicontinuous globally subanalytic function and $f(x_0) = 0$. There exist $\rho > 0$ and a continuous definable function $\chi : \mathbb{R}_+ \to (0,+\infty)$ such that*

$$|f(x)|^\theta \le \rho\,||x^*||,$$

whenever $0 < |f(x)| \le \chi(||x||)$ and $x^* \in \partial^\circ f(x)$.

Proof. In the case that f is globally subanalytic, one can apply [20, Theorem LI] to deduce that the continuous function ψ of Theorem 14 can be taken of the form $\psi(s) = s^{1-\theta}$, with $\theta \in (0,1)$. □

Remark 8. Corollary 9(ii) (and a fortiori Corollary 16) extends [3, Theorem 7] to the lower semicontinuous case. We also remark that the conclusions of Theorem 14 and of Corollary 16 remain valid for any notion of subdifferential that is included in the Clarke subdifferential and thus, in particular, in view of (7), for the Fréchet and the limiting subdifferential. However, let us point out that this is not the case

for broader notions of subdifferentials, as, for example, the *convex-stable* subdifferential introduced and studied in [4]. It is known that the convex-stable subdifferential coincides with the Clarke subdifferential whenever the function f is locally Lipschitz continuous, but it is strictly larger in general, creating more critical points. In particular, [3, section 4] constructs an example of a subanalytic continuous function on $\mathbb{R}^3$ that is strictly increasing in a segment lying in the set of its broadly critical points (that is, critical in the sense of the convex-stable subdifferential). Consequently, Theorem 14 and Corollary 16 do not hold for this subdifferential.

Acknowledgments. A part of this work has been done during a visit of the first author at the Autonomous University of Barcelona (July 2005) and at the University of Nagoya (June 2006). The first author thanks J.-P. Dedieu and the second author K. Kurdyka, S. Simon, and T. Lachand-Robert for useful discussions. The first author thanks the C.R.M. (Barcelona) and the University of Nagoya for financial support.

REFERENCES

[1] P. A. Absil, R. Mahony, and B. Andrews, *Convergence of the iterates of descent methods for analytic cost functions*, SIAM J. Optim., 16 (2005), pp. 531–547.

[2] J. Bolte, A. Daniilidis, and A. Lewis, *Tame functions are semismooth*, Math. Program., (to appear).

[3] J. Bolte, A. Daniilidis, A. Lewis, and M. Shiota, *Clarke critical values of subanalytic Lipschitz continuous functions*, Ann. Polon. Math., 87 (2005), pp. 13–25 (volume dedicated to the memory of S. Łojasiewicz).

[4] J. V. Burke, A. S. Lewis, and M. L. Overton, *A robust gradient sampling algorithm for nonsmooth, nonconvex optimization*, SIAM J. Optim., 15 (2005), pp. 751–779.

[5] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley and Sons, New York, 1983 (republished in Classics Appl. Math. 5, SIAM, 1990, p. 308).

[6] F. H. Clarke, Y. Ledyaev, R. Stern, and P. Wolenski, *Nonsmooth Analysis and Control Theory*, Grad. Texts in Math. 178, Springer, New York, 1998.

[7] M. Coste, *An Introduction to o-minimal Geometry*, RAAG Notes, Institut de Recherche Mathématiques de Rennes, 1999.

[8] D. D'Acunto, *Sur les courbes intégrales du champs de gradient*, Thèse de Doctorat, Université de Savoie, Chambery, France, 2001.

[9] A. Daniilidis, W. Hare, and J. Malick, *Geometrical interpretation of the predictor-corrector type algorithms in structured optimization problems*, Optimization, 55 (2006), pp. 481–503.

[10] Z. Denkowska and K. Wachta, *Une construction de la stratification sous-analytique avec la condition* (w), Bull. Pol. Acad. Sci. Math., 35 (1987), pp. 401–405.

[11] L. van den Dries and C. Miller, *Geometric categories and o-minimal structures*, Duke Math. J., 84 (1996), pp. 497–540.

[12] L. M. Graña Drummond and Y. Peterzil, *The central path in smooth convex semidefinite programs*, Optimization, 51 (2002), pp. 207–233.

[13] A. Ioffe, *A Sard Theorem for Tame Nonsmooth Functions*, J. Math. Anal. Appl., to appear.

[14] A. Ioffe, *Tame optimization: State of the art and perspectives*, Plenary talk at The International Conference on Nonlinear Programming with Applications, Shanghai, 2006.

[15] A. Ioffe, *Critical values of set-valued maps with stratifiable graphs*, in Extensions of Sard and Smale-Sard Theorems, Proc. Amer. Math. Soc., to appear.

[16] A. Ioffe, *Approximate subdifferentials and applications* II, Mathematika, 33 (1986), pp. 111–128.

[17] A. Ioffe, *Approximate subdifferentials and applications* III. *The metric theory*, Mathematika, 36 (1989), pp. 1–38.

[18] V. Kaloshin, *A geometric proof of the existence of Whitney stratifications*, Mosc. Math. J., 5 (2005), pp. 125–133.

[19] K. C. Kiwiel, *Methods of Descent for Nondifferentiable Optimization*, Lecture Notes in Math. 1133, Springer-Verlag, Berlin, 1985.

[20] K. Kurdyka, *On gradients of functions definable in o-minimal structures*, Ann. Inst. Fourier (Grenoble), 48 (1998), pp. 769–783.

[21] K. Kurdyka and A. Parusinski, *w_f-stratification of subanalytic functions and the Łojasiewicz inequality*, C. R. Math. Acad. Sci. Paris, 318 (1994), pp. 129–133.

[22] K. Kurdyka, P. Orro, and S. Simon, *Semialgebraic Sard theorem for generalized critical values*, J. Differential Geom., 56 (2000), pp. 67–92.
[23] C. Lemarechal, F. Oustry, and C. Sagastizabal, *The U-Lagrangian of a convex function*, Trans. Amer. Math. Soc., 352 (2000), pp. 711–729.
[24] A. S. Lewis, *Active sets, nonsmoothness, and sensitivity*, SIAM J. Optim., 13 (2003), pp. 702–725.
[25] S. Łojasiewicz, *Une propriété topologique des sous-ensembles analytiques réels*, in Les Équations aux Dérivées Partielles, Éditions du Centre National de la Recherche Scientifique, Paris, 1963, pp. 87–89.
[26] J. Malick and S. Miller, *Newton methods for nonsmooth convex minimization: Connection among U-Lagrangian, Riemannian Newton and SQP methods*, Math. Program., 104 (2005), pp. 609–633.
[27] J. Mather, *Notes in Topological Stability*, lecture notes, Harvard University, 1970.
[28] B. Mordukhovich, *Variational Analysis and Generalized Differentiation. Basic Theory*, Vol. I, Grundlehren Math. Wiss. 330, Springer-Verlag, Berlin, 2006.
[29] B. Mordukhovich, *Variational Analysis and Generalized Differentiation. Applications*, Vol. II, Grundlehren Math. Wiss. 331, Springer-Verlag, Berlin, 2006.
[30] B. Mordukhovich and Y. Shao, *Nonsmooth sequential analysis in Asplund spaces*, Trans. Amer. Math. Soc., 348 (1996), pp. 1235–1280.
[31] J. Palis and W. De Melo, *Geometric Theory of Dynamical Systems. An Introduction*, Springer-Verlag, New York, 1982 (translated from Portuguese by A. K. Manning).
[32] R. T. Rockafellar and R. Wets, *Variational Analysis*, Grundlehren Math. Wiss. 317, Springer-Verlag, Berlin, 1998.
[33] A. Sard, *The measure of the critical values of differentiable maps*, Bull. Amer. Math. Soc., 48 (1942), pp. 883–890.
[34] M. Shiota, *Geometry of Subanalytic and Semialgebraic Sets*, Progr. Math. 150, Birkhäuser, Boston, 1997.
[35] Ta Lê Loi, *Verdier and strict Thom stratifications in o-minimal structures*, Illinois J. Math., 42 (1998), pp. 347–356.

SIAM J. OPTIM.
Vol. 18, No. 2, pp. 573–588

WEAK SHARP MINIMA FOR SEMI-INFINITE OPTIMIZATION PROBLEMS WITH APPLICATIONS*

XI YIN ZHENG† AND XIAO QI YANG‡

Abstract. We study local weak sharp minima and sharp minima for smooth semi-infinite optimization problems SIP. We provide several dual and primal characterizations for a point to be a sharp minimum or a weak sharp minimum of SIP. As applications, we present several sufficient and necessary conditions of calmness for infinitely many smooth inequalities. In particular, we improve some calmness results in [R. Henrion and J. Outrata, *Math. Program.*, 104 (2005), pp. 437–464].

Key words. semi-infinite optimization, sharp minima, weak sharp minima, subdifferential, normal cone

AMS subject classifications. 90C30, 90C34, 49J52, 65K10

DOI. 10.1137/060670213

1. Introduction.

The notion of a sharp minimum, namely, a strong isolated minimum or a strong unique local minimum, of real-valued functions, introduced in [24], plays an important role in the convergence analysis of numerical algorithms in mathematical programming problems (see [4, 12, 22, 30]). As such, it has received extensive attention and investigation. As a generalization of sharp minima, weak sharp minima for real-valued functions were introduced and studied in [5]. Extensive study of weak sharp minima for real-valued convex functions has been done in the literature (cf. [2, 3, 28, 31, 33]). It has been found that the weak sharp minimum is closely related to the error bound in convex programming (cf. [32]), a notion that has received much attention and has produced a vast number of publications (see [16, 17, 23, 31, 32]).

The calmness is an important type of Lipschitz-like property for multifunctions, which play a key role in many issues of mathematical programming such as sensitivity analysis, error bounds, and optimality conditions. Thus, the study of the calmness has recently received increasing attention in the mathematical programming literature (see [8, 9, 10, 15]).

In this paper, we will study local weak sharp minima for the following semi-infinite optimization problem:

$$\text{(SIP)} \qquad \min f(x) \ \text{ subject to } \phi(x,y) \le 0 \text{ for all } y \in Y,$$

where $f : X \to R$ is a smooth function, X is an Euclidean space, Y is an infinite index set, and $\phi : X \times Y \to R$ is a function such that the function $x \mapsto \phi(x,y)$ is smooth for each index $y \in Y$. It is known that (SIP) has many important and interesting applications in engineering design, control of robots, mechanical stress of

*Received by the editors September 19, 2006; accepted for publication (in revised form) February 26, 2007; published electronically June 12, 2007. This research was supported by the National Natural Science Foundation of People's Republic of China (grant 10361008) and the Natural Science Foundation of Yunnan Province, People's Republic of China (grant 2003A002M).

http://www.siam.org/journals/siopt/18-2/67021.html

†Department of Mathematics, Yunnan University, Kunming 650091, People's Republic of China (xyzheng@ynu.edu.cn).

‡Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hum, Kowloon, Hong Kong (mayangxq@polyu.edu.hk). This author was supported by the Research Grants Council of Hong Kong (PolyU5303/05E).

materials, and social sciences; see the survey paper [11] and the books [6, 21, 25]. In the past three decades, (SIP) and its broad range of applications have been an active study area in mathematical programming (see [1, 7, 13, 14, 20, 27, 29] and references therein).

Let Z denote the set of all feasible points for (SIP); that is,

$$Z := \{x \in X : \ \phi(x, y) \leq 0 \text{ for all } y \in Y\}.$$

We say that $\bar{x} \in X$ is a local sharp minimum of (SIP) if $\bar{x} \in Z$ and there exist $\eta, \delta \in (0, \ +\infty)$ such that

$$\eta\|x - \bar{x}\| \leq f(x) - f(\bar{x}) + \sup_{y \in Y}[\phi(x, y)]_+ \ \text{ for all } x \in B(\bar{x}, \delta), \tag{1.1}$$

where $B(\bar{x}, \delta)$ denotes the open ball with center $\bar{x}$ and radius δ.

We say that $\bar{x}$ is a local weak sharp minimum of (SIP) if $\bar{x} \in Z$ and there exist $\eta, \delta \in (0, \ +\infty)$ such that

$$\eta d(x, L_f(\bar{x}) \cap Z) \leq f(x) - f(\bar{x}) + \sup_{y \in Y}[\phi(x, y)]_+ \ \text{ for all } x \in B(\bar{x}, \delta), \tag{1.2}$$

where $L_f(\bar{x}) := \{x \in X : \ f(x) = f(\bar{x})\}$ and $d(x, L_f(\bar{x}) \cap Z) := \inf\{\|x - u\| : \ u \in L_f(\bar{x}) \cap Z\}$.

Recall a known optimality condition of (SIP) (cf. [11, 13, 34]) that if $\bar{x}$ is a local minimum of (SIP) and a constraint qualification is satisfied at $\bar{x}$, then there exist $t_i \geq 0$ and $y_i \in I_0(\bar{x})$, $i = 1, \ldots, p$, such that

$$0 = f'(\bar{x}) + \sum_{i=1}^{p} t_i \phi'_x(\bar{x}, y_i), \tag{1.3}$$

where $I_0(\bar{x})$ denotes the index set of active inequality constraints at $\bar{x}$. Furthermore, under a convexity assumption, the optimality condition (1.3) also becomes sufficient.

When Y is a compact topological space and $\phi(x, y)$ and $\phi'_x(x, y)$ satisfy some continuity conditions, we will prove that $\bar{x}$ is a local weak sharp minimum of (SIP) if and only if there exist $\eta, \delta \in (0, \ +\infty)$ such that for all $u \in L_f(\bar{x}) \cap Z \cap B(\bar{x}, \delta)$

$$\tilde{N}(L_f(\bar{x}) \cap Z, u) \cap \eta B_{X^*} \subset f'(u) + [0, \ 1]\text{co}\{\phi'_x(u, y) : \ y \in I_0(\bar{x})\}, \tag{1.4}$$

where X^* denotes the dual space of X, B_{X^*} denotes the unit ball of X^*, and $\tilde{N}(A, u)$ is any one of the Fréchet, limiting, or Clarke normal cones of A at u; in particular, $\bar{x}$ is a local sharp minimum of (SIP) if and only if

$$0 \in \text{int}(f'(\bar{x}) + [0, \ 1]\text{co}\{\phi'_x(\bar{x}, y) : \ y \in I_0(\bar{x})\}). \tag{1.5}$$

It is interesting to compare (1.5) and (1.4) with (1.3). These are referred to as dual characterizations. We also obtain a set of primal ones for a local weak sharp minimum of (SIP). Moreover, we obtain mixed characterizations for a local (weak) sharp minimum.

Motivated by Henrion and Outrata [10], we consider the calmness of multifunctions defined by infinitely many smooth inequalities. As applications of several characterizations of weak sharp minima mentioned above, we provide several equivalent conditions for the calmness; in particular, we improve one of the main results in [10].

The outline of the paper is as follows. In section 2, some preliminaries on notions of variational analysis are given. In section 3, several characterizations for a local weak sharp minimum and a local sharp minimum of (SIP) are obtained. In section 4, some equivalent conditions for the calmness of the system of infinitely many smooth inequalities are provided.

2. Preliminaries. Let X be an Euclidean space and $\psi : X \to R \cup \{+\infty\}$ a proper lower semicontinuous function. For $x \in \mathrm{dom}(\psi) := \{x \in X : \ \phi(x) < +\infty\}$, let $\hat{\partial}\psi(x)$ denote the Fréchet subdifferential of ψ at x; that is,

$$\hat{\partial}\psi(x) := \left\{ x^* \in X^* : \ \liminf_{u \overset{\psi}{\to} x} \frac{\psi(u) - \psi(x) - \langle x^*, u - x\rangle}{\|u - x\|} \geq 0 \right\},$$

where $u \overset{\psi}{\to} x$ means $u \to x$ and $\psi(u) \to \psi(x)$. The limiting subdifferential of ψ at x is denoted by $\partial\psi(x)$ and is defined by

$$\partial\psi(x) := \limsup_{u \overset{\psi}{\to} x} \hat{\partial}\psi(u);$$

that is, $x^* \in \partial\psi(x)$ if and only if there exist sequences $x_k \overset{\psi}{\to} x$ and $x_k^* \to x^*$ with $x_k^* \in \hat{\partial}\psi(x_k)$.

The following proposition is well known (cf. [18, Theorem 2.33]) and is useful for us.

PROPOSITION 2.1. *Let $\psi_1, \psi_2 : X \to R \cup \{+\infty\}$ be proper lower semicontinuous functions and $x \in \mathrm{dom}(\psi_1) \cap \mathrm{dom}(\psi_2)$. Suppose that ψ_1 is locally Lipschitz at x. Then*

$$\partial(\psi_1 + \psi_2)(x) \subset \partial\psi_1(x) + \partial\psi_2(x).$$

For a closed subset A of X and $a \in A$, let $\hat{N}(A, a)$ and $N(A, a)$ denote the Fréchet normal cone and the limiting normal cone of A at a, respectively; that is,

$$\hat{N}(A, a) = \hat{\partial}\delta_A(a) \ \text{ and } \ N(A, a) = \partial\delta_A(a),$$

where δ_A denotes the indicator function of A. Thus, $x^* \in \hat{N}(A, a)$ if and only if $\limsup_{x \overset{A}{\to} a} \frac{\langle x^*, x-a\rangle}{\|x-a\|} \leq 0$, where $x \overset{A}{\to} a$ means $x \in A$ and $x \to a$, and $x^* \in N(A, a)$ if and only if there exist $x_k \overset{A}{\to} a$ and $x_k^* \to x^*$ such that $x_k^* \in \hat{N}(A, x_k)$ for all $k \in \mathbb{N}$, where $\mathbb{N}$ denotes the set of all natural numbers.

Let $T(A, a)$ denote the tangent cone of A at a; that is,

$$T(A, a) := \{h \in X : \ \exists t_k \to 0^+ \text{ and } h_k \to h \text{ such that } a + t_k h_k \in A \text{ for all } k \in \mathbb{N}\}.$$

It is known (cf. [26, Theorem 6.28]) that

$$\hat{N}(A, a) = \{x^* \in X^* : \ \langle x^*, h\rangle \leq 0 \text{ for all } h \in T(A, a)\}. \tag{2.1}$$

Let $T_c(A, a)$ denote the Clarke tangent cone; that is, $v \in T_c(A, a)$ if and only if, for each sequence $\{a_k\}$ in A converging to a and each sequence $\{t_k\}$ in $(0, \infty)$ decreasing to 0, there exists a sequence $\{v_k\}$ in X converging to v such that $a_k + t_k v_k \in A$ for all $k \in \mathbb{N}$. Let $N_c(A, a)$ denote the Clarke normal cone of A at a and be defined by

$$N_c(A, a) := \{x^* \in X^* : \ \langle x^*, v\rangle \leq 0 \quad \text{for all } v \in T_c(A, a)\}. \tag{2.2}$$

It is well known (cf. [26, Proposition 6.5] and [18, Theorem 3.57]) that

$$\hat{N}(A,a) \subset N(A,a) \subset N_c(A,a) \text{ and } N_c(A,a) = \overline{\text{co}}N(A,a). \tag{2.3}$$

Histories of the subdifferentials and the normal cones can be found in [18, 19, 26].

For any $x \in X$, let $P_A(x)$ denote the projection of x on A; that is,

$$P_A(x) := \{a \in A : \ \|x-a\| = d(x,A)\}.$$

We will need the following known result (cf. [26, Example 6.16]).

LEMMA 2.1. *Let A be a closed subset of X and $x \in X$. Then*

$$x - a \in \hat{N}(A,a) \text{ for any } a \in P_A(x). \tag{2.4}$$

3. Weak sharp minima for smooth semi-infinite optimization problems. Throughout the remainder of this paper, let X be an Euclidean space of dimension m and Y a compact topological space (e.g., a bounded closed subset of an Euclidean space). Let $f : X \to R$ and $\phi : X \times Y \to R$ be as in section 1. We always assume that the following properties hold:

(P1) The function $x \mapsto \phi(x,y)$ is smooth for each $y \in Y$, and the function $y \mapsto \phi(x,y)$ is continuous for each $x \in X$.

(P2) The functions $(x,y) \mapsto \phi(x,y)$ and $(x,y) \mapsto \phi'_x(x,y)$ are continuous on $X \times Y$, where $\phi'_x(x,y)$ denotes the derivative of the function $x \mapsto \phi(x,y)$.

In the literature on semi-infinite optimization, assumptions (P1) and (P2) have been extensively used.

Since Y is compact and (P1) holds, it is easy to verify that $\bar{x} \in Z$ is a local sharp minimum and a local weak sharp minimum of (SIP) if and only if there exist $\eta, \delta \in (0, \ +\infty)$ such that

$$\eta\|x-\bar{x}\| \le f(x) - f(\bar{x}) + \max_{y\in Y}[\phi(x,y)]_+ \text{ for all } x \in B(\bar{x},\delta) \tag{3.1}$$

and

$$\eta d(x, L_f(\bar{x}) \cap Z) \le f(x) - f(\bar{x}) + \max_{y\in Y}[\phi(x,y)]_+ \text{ for all } x \in B(\bar{x},\delta), \tag{3.2}$$

respectively.

It follows from (3.2) that every local weak sharp minimum of (SIP) is a local solution of (SIP). Clearly, $\bar{x}$ is a local sharp minimum of (SIP) if and only if $\bar{x}$ is a local weak sharp minimum of (SIP) and

$$L_f(\bar{x}) \cap Z \cap B(\bar{x},\delta) = \{\bar{x}\} \text{ for some } \delta > 0.$$

For convenience, let

$$\Phi(x) := \max\{\phi(x,y) : \ y \in Y\} \text{ and } I(x) := \{y \in Y : \ \phi(x,y) = \Phi(x)\}.$$

From (P1) and the compactness of Y, it is clear that $I(x) \neq \emptyset$ for all $x \in X$. For each $x \in Z$, let $I_0(x)$ denote the index set of active inequality constraints at x; that is,

$$I_0(x) := \{y \in Y : \ \phi(x,y) = 0\}.$$

We will provide characterizations for $\bar{x}$ to be a local weak sharp minimum or a local sharp minimum of (SIP). We need the following lemma.

LEMMA 3.1. *Let $\bar{x} \in X$ and $\varepsilon > 0$. Then there exists $\delta > 0$ such that for any $x \in B(\bar{x}, \delta)$ and $u \in L_f(\bar{x}) \cap Z \cap B(\bar{x}, \delta)$*

$$\langle f'(u), x - u\rangle \le f(x) - f(\bar{x}) + \varepsilon\|x - u\|$$

and

$$\langle \phi'_x(u, y), x - u\rangle \le \phi(x, y) + \varepsilon\|x - u\| \quad \textit{for all } y \in I_0(u).$$

Proof. Since $(x, y) \mapsto \phi'_x(x, y)$ is continuous, for any $y \in Y$ there exist open neighborhoods U_y and V_y of $\bar{x}$ and y, respectively, such that

$$\|\phi'_x(x_1, v_1) - \phi'_x(x_2, v_2)\| < \varepsilon \quad \text{for all } x_1, x_2 \in U_y \text{ and for all } v_1, v_2 \in V_y.$$

Since Y is compact, there exist $y_1, \ldots, y_k \in Y$ such that $Y = \bigcup_{i=1}^{k} V_{y_i}$. Let $U := \bigcap_{i=1}^{n} U_{y_i}$, and take $\delta > 0$ such that $B(\bar{x}, \delta) \subset U$. It is easy to verify that

$$\|\phi'_x(x_1, y) - \phi'_x(x_2, y)\| < \varepsilon \quad \text{for all } x_1, x_2 \in B(\bar{x}, \delta) \text{ and for all } y \in Y. \tag{3.3}$$

Since f is continuously differentiable, we assume without loss of generality that

$$\|f'(x_1) - f'(x_2)\| < \varepsilon \quad \text{for all } x_1, x_2 \in B(\bar{x}, \delta) \tag{3.4}$$

(considering smaller δ if necessary). Let $x \in B(\bar{x}, \delta)$, $u \in L_f(\bar{x}) \cap Z \cap B(\bar{x}, \delta)$, and $y \in I_0(u)$. By the mean value theorem, there exist $\theta_1, \theta_2 \in (u,\ x) := \{tu + (1 - t)x : 0 < t < 1\}$ such that

$$f(x) - f(\bar{x}) = f(x) - f(u) = \langle f'(\theta_1), x - u\rangle$$

and

$$\phi(x, y) = \phi(x, y) - \phi(u, y) = \langle \phi'_x(\theta_2, y), x - u\rangle.$$

It follows from (3.4) and (3.3) that

$$\begin{aligned}\langle f'(u), x - u\rangle &= \langle f'(u) - f'(\theta_1), x - u\rangle + \langle f'(\theta_1), x - u\rangle \\ &\le f(x) - f(\bar{x}) + \varepsilon\|x - u\|\end{aligned}$$

and

$$\langle \phi'_x(u, y), x - u\rangle \le \phi(x, y) + \varepsilon\|x - u\|.$$

The proof is completed. □

LEMMA 3.2. *Let $\bar{x} \in Z$ and $u \in L_f(\bar{x}) \cap Z$. Then*

$$\langle f'(u), h\rangle = 0 \ \text{ and } \ \langle \phi'_x(u, y), h\rangle \le 0, \quad \textit{for all } h \in T(L_f(\bar{x}) \cap Z, u) \textit{ and for all } y \in I_0(u).$$

Proof. Let $h \in T(L_f(\bar{x}) \cap Z, u)$ and $y \in I_0(u)$. Then there exist $t_k \to 0^+$ and $h_k \to h$ such that $u + t_k h_k \in L_f(\bar{x}) \cap Z$ for all $k \in \mathbb{N}$. Hence

$$f(u + t_k h_k) = f(u) = f(\bar{x}) \ \text{ and } \ \phi(u + t_k h_k, y) \le 0 \quad \text{for all } k \in \mathbb{N}.$$

Since f is continuously differentiable,

$$f(u + t_k h_k) - f(u) = \langle f'(u), t_k h_k\rangle + o(t_k).$$

It follows that $\langle f'(u), h_k\rangle + \frac{o(t_k)}{t_k} = 0$. This implies that $\langle f'(u), h\rangle = 0$. Let ε be an arbitrary positive number. Then Lemma 3.1 implies that

$$\langle \phi'(u,y), t_k h_k\rangle \le \phi(u + t_k h_k, y) + \varepsilon\|t_k h_k\| \le \varepsilon\|t_k h_k\|$$

for all k large enough, and so $\langle \phi'(u,y), h\rangle \le \varepsilon\|h\|$. Since ε is arbitrary, it follows that $\langle \phi'(u,y), h\rangle \le 0$. This completes the proof. □

In the next theorems we first provide some dual characterizations and then some primal characterizations for a feasible point of (SIP) to be a local weak sharp minimum. As usual, let co$A$ denote the convex hull of A. For convenience, we adopt the conventions that if $u \in L_f(\bar{x}) \cap Z$, with $I_0(u) = \emptyset$, then

$$[0,\ 1]\mathrm{co}\{\phi'_x(u,y):\ y \in I_0(u)\} := \{0\}$$

and

$$\max_{y\in I_0(u)} [\langle \phi'_x(u,y), h\rangle]_+ := 0 \ \text{ for all } h \in X.$$

THEOREM 3.1. *Let $\bar{x}$ be a feasible point of* (SIP) *(i.e., $\bar{x} \in Z$). Then the following statements are equivalent:*

(i) *$\bar{x}$ is a local weak sharp minimum of* (SIP).

(ii) *There exist $\eta, \delta \in (0,\ +\infty)$ such that*

$$\hat{N}(L_f(\bar{x}) \cap Z, u) \cap \eta B_{X^*} \subset f'(u) + [0,\ 1]\mathrm{co}\{\phi'_x(u,y):\ y \in I_0(u)\} \tag{3.5}$$

for all $u \in L_f(\bar{x}) \cap Z \cap B(\bar{x}, \delta)$.

(iii) *There exist $\eta, \delta \in (0,\ +\infty)$ such that*

$$N(L_f(\bar{x}) \cap Z, u) \cap \eta B_{X^*} \subset f'(u) + [0,\ 1]\mathrm{co}\{\phi'_x(u,y):\ y \in I_0(u)\} \tag{3.6}$$

for all $u \in L_f(\bar{x}) \cap Z \cap B(\bar{x}, \delta)$.

(iv) *There exist $\eta, \delta \in (0,\ +\infty)$ such that*

$$N_c(L_f(\bar{x}) \cap Z, u) \cap \eta B_{X^*} \subset f'(u) + [0,\ 1]\mathrm{co}\{\phi'_x(u,y):\ y \in I_0(u)\} \tag{3.7}$$

for all $u \in L_f(\bar{x}) \cap Z \cap B(\bar{x}, \delta)$.

Proof. (i)⇒(ii). Suppose that there exist $\eta, \delta \in (0,\ +\infty)$ such that (3.2) holds. Let $u \in L_f(\bar{x}) \cap Z \cap B(\bar{x}, \frac{\delta}{2})$ and $u^* \in \hat{N}(L_f(\bar{x}) \cap Z, u) \cap B_{X^*}$. Let $\varepsilon > 0$, and take $r \in (0,\ \frac{\delta}{2})$ such that

$$\langle u^*, v - u\rangle \le \varepsilon\|v - u\| \quad \text{for all } v \in L_f(\bar{x}) \cap Z \cap B(u, r).$$

Let $x \in B(u, \frac{r}{2}) \subset B(\bar{x}, \delta)$. Then there exists $v \in L_f(\bar{x}) \cap Z$ such that $\|x - v\| = d(x, L_f(\bar{x}) \cap Z)$. Hence,

$$\|v - u\| \le \|v - x\| + \|x - u\| \le 2\|x - u\| < r.$$

Therefore,

$$\begin{aligned}\langle u^*, x - u\rangle &= \langle u^*, x - v\rangle + \langle u^*, v - u\rangle \\ &\le \|x - v\| + \varepsilon\|v - u\| \\ &\le (1 + \varepsilon)\|x - v\| + \varepsilon\|x - u\| \\ &= (1 + \varepsilon)d(x, L_f(\bar{x}) \cap Z) + \varepsilon\|x - u\|.\end{aligned}$$

It follows from (3.2) and $B(u, \frac{r}{2}) \subset B(\bar{x}, \delta)$ that

$$\eta\langle u^*, x-u\rangle \le (1+\varepsilon)(f(x)-f(\bar{x})+[\Phi(x)]_+)+\eta\varepsilon\|x-u\| \quad \text{for all } x \in B\left(u, \frac{r}{2}\right).$$

Noting that $f(u) = f(\bar{x})$ and $[\Phi(u)]_+ = 0$, it follows that u is a local minimum of the function

$$x \mapsto -\eta\langle u^*, x-u\rangle + (1+\varepsilon)(f(x)-f(\bar{x})+[\Phi(x)]_+)+\eta\varepsilon\|x-u\|.$$

This and Proposition 2.1 imply that

$$\eta u^* \in (1+\varepsilon)(f'(u)+\partial[\Phi(\cdot)]_+(u))+\eta\varepsilon B_{X^*}.$$

Letting $\varepsilon \to 0$, one has

$$\eta u^* \in f'(u)+\partial[\Phi(\cdot)]_+(u) = f'(u)+\mathrm{co}(\{0\}\cup\partial\Phi(u)) = f'(u)+[0,\ 1]\partial\Phi(u).$$

Noting (by [26, Theorem 10.31]) that

$$\partial\Phi(u) = \begin{cases} \{0\} & I_0(u)=\emptyset, \\ \mathrm{co}\{\phi'_x(u,y):\ y\in I_0(u)\} & I_0(u)\neq\emptyset, \end{cases}$$

it follows that (3.5) holds.

(ii)⇒(iii). Suppose that there exist $\eta, \delta \in (0,\ +\infty)$ such that (3.5) holds for all $u \in L_f(\bar{x})\cap Z\cap B(\bar{x},\delta)$. Let $u \in L_f(\bar{x})\cap Z\cap B(\bar{x},\delta)$ and $u^* \in N(L_f(\bar{z})\cap Z, u)\cap\eta B_{X^*}$. Take a sequence $\{u_k\}$ in $L_f(\bar{x})\cap Z$ and a sequence $\{u_k^*\}$ in X^* such that $u_k \to u$, $u_k^* \to u^*$, and $u_k^* \in \hat{N}(L_f(\bar{x})\cap Z, u_k)$ for all $k \in \mathbb{N}$. Without loss of generality, we assume that $u_k \in B(\bar{x},\delta)$ and $u_k^* \in \eta B_{X^*}$ for each $k \in \mathbb{N}$. By (3.5), one has

$$u_k^* \in f'(u_k)+[0,\ 1]\mathrm{co}\{\phi'_x(u_k,y):\ y\in I_0(u_k)\} \quad \text{for all } k\in\mathbb{N}.$$

We divide into two cases: 1) $I_0(u_k)=\emptyset$ for infinitely many k and 2) $I_0(u_k)\neq\emptyset$ for infinitely many k.

Case 1. Without loss of generality we assume that $I_0(u_k)=\emptyset$ for all $k\in\mathbb{N}$ (passing to a subsequence if necessary). Thus, $u_k^* = f'(u_k)$ for all $k\in\mathbb{N}$. It follows that $u^* = f'(u)$. Hence (3.6) holds.

Case 2. We can assume that $I_0(u_k)\neq\emptyset$ for all $k\in\mathbb{N}$. Noting that X is of dimension m, it follows from the Caratheodory theorem (cf. [26, Theorem 2.29]) that there exist $t_{1k},\dots,t_{m+1k}\in[0,\ 1]$ and $y_{1k},\dots,y_{m+1k}\in I_0(u_k)$ such that

$$\sum_{i=1}^{m+1} t_{ik} \le 1 \ \text{ and } \ u_k^* = f'(u_k)+\sum_{i=1}^{m+1} t_{ik}\phi'_x(u_k,y_{ik}) \quad \text{for all } k\in\mathbb{N}.$$

Without loss of generality, we assume that

$$t_{ik}\to t_i \ \text{ and } \ y_{ik}\to y_i\in I_0(u) \ \text{ as } k\to\infty,\ i=1,\dots,m+1$$

(passing to subsequences if necessary). Thus,

$$\sum_{i=1}^{m+1} t_i \le 1 \quad \text{and} \quad u^* = f'(u)+\sum_{i=1}^{m+1} t_i\phi'_x(u,y_i).$$

This shows that (3.6) holds for all $u\in L_f(\bar{x})\cap Z\cap B(\bar{x},\delta)$.

(iii)⇒(iv). Suppose that there exist $\eta, \delta \in (0, +\infty)$ such that (3.6) holds for all $u \in L_f(\bar{x}) \cap Z \cap B(\bar{x}, \delta)$. It follows that

$$N(L_f(\bar{x})\cap Z, u) \subset R_+ f'(u)+R_+\mathrm{co}\{\phi'_x(u,y) :\ y \in I_0(u)\} \quad \text{for all } u \in L_f(\bar{x})\cap Z\cap B(\bar{x},\delta).$$

On the other hand, by (2.1) and Lemma 3.2 one has

$$R_+ f'(u) + R_+\mathrm{co}\{\phi'_x(u,y) :\ y \in I_0(u)\} \subset \hat{N}(L_f(\bar{x}) \cap Z, u) \quad \text{for all } u \in L_f(\bar{x}) \cap Z.$$

It follows that

$$\hat{N}(L_f(\bar{x}) \cap Z, u) = N(L_f(\bar{x}) \cap Z, u) \quad \text{for all } u \in L_f(\bar{x}) \cap Z \cap B(\bar{x}, \delta).$$

By (2.1) and (2.3), one has

$$N(L_f(\bar{x}) \cap Z, u) = N_c(L_f(\bar{x}) \cap Z, u) \quad \text{for all } u \in L_f(\bar{x}) \cap Z \cap B(\bar{x}, \delta).$$

Therefore, (iv) holds.

(iv)⇒(i). Suppose that there exist $\eta, \delta \in (0, +\infty)$ such that (3.7) holds for all $x \in L_f(\bar{x})\cap Z\cap B(\bar{x}, \delta)$. Let $x \in B(\bar{x}, \frac{\delta}{2}) \setminus L_f(\bar{x})\cap Z$, and take $u \in P_{L_f(\bar{x})\cap Z}(x)$. Then $u \in B(\bar{x}, \delta)$, and it follows from Lemma 2.1 and (2.3) that $\frac{x-u}{\|x-u\|} \in N_c(L_f(\bar{x}) \cap Z, u)$. We claim that $I_0(u) \neq \emptyset$. Suppose to the contrary that $I_0(u) = \emptyset$. Then, by the definition, $[0,\ 1]\mathrm{co}\{\phi'_x(u,y) :\ u \in I_0(u)\} = \{0\}$. This and (3.7) imply that the intersection $N_c(L_f(\bar{x}) \cap Z, u) \cap \eta B_{X^*}$ is the singleton $\{f'(u)\}$, contradicting the fact that it contains 0 and $\frac{\eta(x-u)}{\|x-u\|}$. Hence $I_0(u) \neq \emptyset$. By (3.7), there exist $t_1, \ldots, t_q \in [0, +\infty)$ and $y_1, \ldots, y_q \in I_0(u)$ such that

$$\sum_{i=1}^{q} t_i \leq 1 \ \text{ and } \ \frac{\eta(x-u)}{\|x-u\|} = f'(u) + \sum_{i=1}^{q} t_i \phi'_x(u, y_i).$$

Hence

$$\eta\|x-u\| = \langle f'(u), x-u\rangle + \sum_{i=1}^{q} t_i \langle \phi'_x(u,y_i), x-u\rangle.$$

Let $\varepsilon \in (0,\ \frac{\eta}{2})$. By Lemma 3.1, without loss of generality we assume that

$$\langle f'(u), x-u\rangle \leq f(x) - f(\bar{x}) + \varepsilon\|x-u\|$$

and

$$\langle \phi'_x(u,y), x-u\rangle \leq \phi(x,y) + \varepsilon\|x-u\| \quad \text{for all } y \in I_0(u)$$

(considering smaller δ if necessary). Therefore,

$$\begin{aligned}\eta\|x-u\| &\leq f(x) - f(\bar{x}) + \sum_{i=1}^{q} t_i \phi(x, y_i) + 2\varepsilon\|x-u\| \\ &\leq f(x) - f(\bar{x}) + \max_{y\in Y}[\phi(x,y)]_+ + 2\varepsilon\|x-u\|.\end{aligned}$$

It follows that

$$(\eta - 2\varepsilon)\|x-u\| \leq f(x) - f(\bar{x}) + \max_{y\in Y}[\phi(x,y)]_+;$$

that is,

$$(\eta - 2\varepsilon) d(x, L_f(\bar{x}) \cap Z) \leq f(x) - f(\bar{x}) + \max_{y\in Y}[\phi(x,y)]_+.$$

Since $f(x) = f(\bar{x})$ and $\max_{y\in Y}[\phi(x,y)]_+ = 0$ if $x \in L_f(\bar{x}) \cap Z$, the last inequality holds trivially if $x \in L_f(\bar{x}) \cap Z$. This shows that (i) holds. The proof is completed. □

Remark. In view of the proof of Theorem 3.1, one can see that the implication (i)⇒(ii) of Theorem 3.1 holds even when X is a Banach space of infinite dimension.

THEOREM 3.2. *Let $\bar{x} \in Z$. Then the following statements are equivalent:*

(i) *$\bar{x}$ is a local weak sharp minimum of* (SIP).

(ii) *There exist $\eta, \delta \in (0, +\infty)$ such that*

$$\eta d(h, T(L_f(\bar{x}) \cap Z, u)) \le \langle f'(u), h\rangle + \max_{y\in I_0(u)}[\langle \phi'_x(u,y), h\rangle]_+ \tag{3.8}$$

for all $u \in L_f(\bar{x}) \cap Z \cap B(\bar{x}, \delta)$ and $h \in X$.

(iii) *There exist $\eta, \delta \in (0, +\infty)$ such that*

$$\eta d(h, T_c(L_f(\bar{x}) \cap Z, u)) \le \langle f'(u), h\rangle + \max_{y\in I_0(u)}[\langle \phi'_x(u,y), h\rangle]_+ \tag{3.9}$$

for all $u \in L_f(\bar{x}) \cap Z \cap B(\bar{x}, \delta)$ and $h \in X$.

(iv) *There exist $\eta, \delta \in (0, +\infty)$ such that*

$$\eta\|x - u\| \le \langle f'(u), x - u\rangle + \max_{y\in I_0(u)}[\langle \phi'_x(u,y), x - u\rangle]_+ \tag{3.10}$$

for any $x \in B(\bar{x}, \delta)$ and $u \in P_{L_f(\bar{x})\cap Z}(x)$.

Proof. (i)⇒(iii). Suppose that (i) holds. Then by Theorem 3.1 there exist $\eta, \delta \in (0, +\infty)$ such that (3.7) holds for all $u \in L_f(\bar{x}) \cap Z \cap B(\bar{x}, \delta)$. Let $u \in L_f(\bar{x}) \cap Z \cap B(\bar{x}, \delta)$ and $h \in X$. By Lemma 3.2, (3.9) holds if $h \in T_c(L_f(\bar{x}) \cap Z, u)$. Now we assume that $h \notin T_c(L_f(\bar{x}) \cap Z, u)$. Take $h_0 \in P_{T_c(L_f(\bar{x})\cap Z,u)}(h)$. Then by Lemma 2.1 and (2.3) one has

$$h - h_0 \in N_c(T_c(L_f(\bar{x}) \cap Z, u), h_0).$$

Since $T_c(L_f(\bar{x}) \cap Z, u)$ is a convex cone,

$$\langle h - h_0, z - h_0\rangle \le 0 \quad \text{for all } z \in T_c(L_f(\bar{x}) \cap Z, u).$$

Hence

$$\langle h - h_0, h_0\rangle = 0 \ \text{ and } \ \langle h - h_0, z\rangle \le 0 \quad \text{for all } z \in T_c(L_f(\bar{x}) \cap Z, u).$$

This and (2.2) imply that $\frac{\eta(h-h_0)}{\|h-h_0\|} \in N_c(L_{f(\bar{x})} \cap Z, u)$. It follows from (3.7) that $I_0(u) \neq \emptyset$, and there exist $t_1, \ldots, t_q \in [0, +\infty)$ and $y_1, \ldots, y_q \in I_0(u)$ such that

$$\sum_{i=1}^{q} t_i \le 1 \ \text{ and } \ \frac{\eta(h - h_0)}{\|h - h_0\|} = f'(u) + \sum_{i=1}^{q} t_i \phi'_x(u, y_i).$$

Hence

$$\begin{aligned}
\eta d(h, T_c(L_f(\bar{x}) \cap Z, u)) &= \left\langle \frac{\eta(h - h_0)}{\|h - h_0\|}, h - h_0 \right\rangle \\
&= \left\langle \frac{\eta(h - h_0)}{\|h - h_0\|}, h \right\rangle \\
&= \langle f'(u), h\rangle + \sum_{i=1}^{q} t_i \langle \phi'_x(u, y_i), h\rangle \\
&\le \langle f'(u), h\rangle + \max_{y\in I_0(u)}[\langle \phi'_x(u,y), h\rangle]_+.
\end{aligned}$$

Therefore, (3.9) holds. This shows that (iii) holds.

Since $T_c(L_f(\bar{x})\cap Z,u)\subset T(L_f(\bar{x})\cap Z,u)$ for any $u\in L_f(\bar{x})\cap Z$,

$$d(h,T(L_f(\bar{x})\cap Z,u))\le d(h,T_c(L_f(\bar{x})\cap Z,u)) \quad \text{for all } h\in X \text{and for all } u\in L_f(\bar{x})\cap Z.$$

Hence (iii)⇒(ii) holds trivially.

Suppose that (ii) holds. Take $\eta,\delta\in(0,\ +\infty)$ such that (3.8) holds for all $u\in L_f(\bar{x})\cap Z\cap B(\bar{x},\delta)$ and $h\in X$. Let $x\in B(\bar{x},\frac{\delta}{2})\setminus L_f(\bar{x})\cap Z$, and take $u\in P_{L_f(\bar{x})\cap Z}(x)$. By Lemma 2.1, one has $\frac{x-u}{\|x-u\|}\in\hat{N}(L_f(\bar{x})\cap Z,u)$. Hence $\langle\frac{x-u}{\|x-u\|},z\rangle\le 0$ for any $z\in T(L_f(\bar{x})\cap Z,u)$. This implies that

$$\|x-u\|\le\left\langle\frac{x-u}{\|x-u\|},x-u-z\right\rangle\le\|x-u-z\| \quad \text{for all } z\in T(L_f(\bar{x})\cap Z,u).$$

Hence $\|x-u\|=d(x-u,T(L_f(\bar{x})\cap Z,u))$. Noting that $u\in B(\bar{x},\delta)$, it follows from (3.8) that (3.10) holds. This shows that the implication (ii)⇒(iv) holds.

Suppose that (iv) holds. Take $\eta,\delta\in(0,\ +\infty)$ such that (3.10) holds for any $x\in B(\bar{x},\delta)$ and $u\in P_{L_f(\bar{x})\cap Z}(x)$. Let $x\in B(\bar{x},\frac{\delta}{2})\setminus L_f(\bar{x})\cap Z$, and take $u\in P_{L_f(\bar{x})\cap Z}(x)$. Then $u\in B(\bar{x},\delta)$. Hence (3.10) holds for such x and u. Let $\varepsilon\in(0,\ \frac{\eta}{2})$. By Lemma 3.1, without loss of generality we assume that

$$\langle f'(u),x-u\rangle\le f(x)-f(\bar{x})+\varepsilon\|x-u\|$$

and

$$\langle\phi'_x(u,y),x-u\rangle\le\phi(x,y)+\varepsilon\|x-u\| \quad \text{for all } y\in I_0(u)$$

(taking smaller δ if necessary). Hence

$$[\langle\phi'_x(u,y),x-u\rangle]_+\le[\phi(x,y)]_+ +\varepsilon\|x-u\| \quad \text{for all } y\in I_0(u).$$

It follows from (3.10) that

$$\eta\|x-u\|\le f(x)-f(\bar{x})+\max_{y\in Y}[\phi(x,y)]_+ +2\varepsilon\|x-u\|.$$

Therefore,

$$(\eta-2\varepsilon)d(x,L_f(\bar{x})\cap Z)=(\eta-2\varepsilon)\|x-u\|\le f(x)-f(\bar{x})+\max_{y\in Y}[\phi(x,y)]_+.$$

This shows that (i) holds. The proof is completed. □

Now we provide a mixed characterization for $\bar{x}$ to be a weak sharp minimum of (SIP), which is inspired from [10, Theorem 4].

PROPOSITION 3.1. *Let $\bar{x}\in Z$. Then $\bar{x}$ is a local weak sharp minimum of* (SIP) *if and only if the following conditions are satisfied:*

(i) $T(L_f(\bar{x})\cap Z,\bar{x})=\{h\in X:\ \langle f'(\bar{x}),h\rangle+\max_{y\in I_0(\bar{x})}[\langle\phi'_x(\bar{x},y),h\rangle]_+\le 0\}$.

(ii) *There exist $\eta_0,\delta\in(0,\ +\infty)$ such that for any $u\in L_f(\bar{x})\cap Z\cap B(\bar{x},\delta)\setminus\{\bar{x}\}$*

$$\hat{N}(L_f(\bar{x})\cap Z,u)\cap\eta_0 B_{X^*}\subset f'(u)+[0,\ 1]\mathrm{co}\{\phi'_x(u,y):\ y\in I_0(u)\}.$$

Proof. By Lemma 3.2, one has

$$T(L_f(\bar{x})\cap Z,\bar{x})\subset\{h\in X:\ \langle f'(\bar{x}),h\rangle+\max_{y\in I_0(\bar{x})}[\langle\phi'_x(\bar{x},y),h\rangle]_+\le 0 \quad \text{for all } y\in I_0(\bar{x})\}.$$

It follows from (ii) of Theorem 3.2 and (ii) of Theorem 3.1 that the necessity part holds.

To prove the sufficiency part, suppose that (i) and (ii) hold. We claim that there exists $\eta_1 > 0$ such that

$$\eta_1\|h\| \le \langle f'(\bar{x}), h\rangle + \max_{y\in I_0(\bar{x})}[\langle \phi'_x(\bar{x}, y), h\rangle]_+ \quad \text{for all } h \in \hat{N}(L_f(\bar{x}) \cap Z, \bar{x}). \tag{3.11}$$

Suppose to the contrary that there exists a sequence $\{h_k\}$ in $\hat{N}(L_f(\bar{x}) \cap Z, \bar{x})$ such that

$$\|h_k\| = 1 \ \text{ and } \ \langle f'(\bar{x}), h_k\rangle + \max_{y\in I_0(\bar{x})}[\langle \phi'_x(\bar{x}, y), h_k\rangle]_+ < \frac{1}{k} \quad \text{for all } k \in \mathbb{N}.$$

Without loss of generality we assume that $h_k \to h_0$. Then

$$h_0 \in \hat{N}(L_f(\bar{x}) \cap Z, \bar{x}) \ \text{ and } \ \langle f'(\bar{x}), h_0\rangle + \max_{y\in I_0(\bar{x})}[\langle \phi'_x(\bar{x}, y), h_0\rangle]_+ \le 0.$$

It follows from (i) that $h_0 \in T(L_f(\bar{x}) \cap Z, \bar{x})$, contradicting $\|h_0\| = 1$ and (2.1). This shows that (3.11) holds. Let $x \in B(\bar{x}, \frac{\delta}{2}) \setminus L_f(\bar{x}) \cap Z$ and $u \in P_{L_f(\bar{x})\cap Z}(x)$. Then $u \in B(\bar{x}, \delta) \setminus \{x\}$ and $\frac{x-u}{\|x-u\|} \in \hat{N}(L_f(\bar{x}) \cap Z, u)$. In the case when $u = \bar{x}$, by (3.11) one has

$$\eta_1\|x - \bar{x}\| \le \langle f'(\bar{x}), x - \bar{x}\rangle + \max_{y\in I_0(\bar{x})}[\langle \phi'_x(\bar{x}, y), x - \bar{x}\rangle]_+. \tag{3.12}$$

In the case when $u \ne \bar{x}$, by (ii) there exist $t_i \in [0, +\infty)$ and $y_i \in I_0(u)$, $i = 1, \ldots, q$, such that

$$\sum_{i=1}^{q} t_i \le 1 \ \text{ and } \ \frac{\eta_0(x-u)}{\|x-u\|} = f'(u) + \sum_{i=1}^{q} t_i\phi'(u, y_i).$$

It follows that

$$\begin{aligned}\eta_0\|x-u\| &= \langle f'(u), x-u\rangle + \sum_{i=1}^{q} t_i\langle \phi'_x(u, y_i), x-u\rangle \\ &\le \langle f'(u), x-u\rangle + \max_{y\in I_0(u)}[\langle \phi'_x(u, y_i), x-u\rangle].\end{aligned}$$

This and (3.12) imply that (iv) of Theorem 3.2 holds with $\eta = \min\{\eta_0, \eta_1\}$. It follows from Theorem 3.2 that the sufficiency part holds. The proof is completed. □

Remark. Letting

$$\psi(x) := f(x) + \max_{y\in Y}[\phi(x, y)]_+ \ \text{ for all } x \in X,$$

it is clear that if $\bar{x}$ is a local weak sharp minimum of (SIP), then $\bar{x}$ is a local weak sharp minimum of ψ: There exist $\eta, \delta \in (0, +\infty)$ such that

$$\eta d(x, L_\psi(\bar{x})) \le \psi(x) - \psi(\bar{x}) \ \text{ for all } x \in B(\bar{x}, \delta).$$

The converse implication may not be true. Indeed, let $X = R$, $Y = \{y_0\}$, $f(x) = -x^2$, and $\phi(x, y_0) = x^2$ for all $x \in R$. Then $Z = \{0\}$, and $\bar{x} = 0$ is not a local weak sharp minimum of (SIP). But, noting that $\psi(x) = f(x) + \max_{y\in Y}[\phi(x, y)]_+ = 0$ for all $x \in X$, 0 is a weak sharp minimum of ψ.

When ψ is a convex function, in terms of the normal and tangent cones of the solution set as well as the subdifferential and the directional derivative of ψ, some characterizations for the weak sharp minimum of ψ have been established (cf. [2, 33]). To the best of our knowledge, in the nonconvex case no one considers corresponding characterizations.

Finally, we provide characterizations for $\bar{x} \in Z$ to be a local sharp minimum of (SIP).

THEOREM 3.3. *Let $\bar{x} \in Z$. Then the following statements are equivalent:*

(i) *$\bar{x}$ is a local sharp minimum of* (SIP).

(ii) *There exists $\eta > 0$ such that*

$$\eta B_{X^*} \subset f'(\bar{x}) + [0,\ 1]\mathrm{co}\{\phi'_x(\bar{x}, y) :\ y \in I_0(\bar{x})\}.$$

(iii) *There exists $\eta > 0$ such that*

$$\eta\|h\| \le \langle f'(\bar{x}), h\rangle + \max_{y\in I_0(\bar{x})}[\langle\phi'_x(\bar{x}, y), h\rangle]_+ \quad \textit{for all } h \in X.$$

(iv) $\{h \in X : \langle f'(\bar{x}), h\rangle + \max_{y\in I_0(\bar{x})}[\langle\phi'_x(\bar{x}, y), h\rangle]_+ \le 0\} = \{0\}$.

Proof. (i)⇒(ii) is immediate from Theorem 3.1 and $N(\{\bar{x}\}, \bar{x}) = X^*$. (ii)⇒(iii) and (iii)⇒(iv) are trivial.

It remains to prove (iv)⇒(i). Suppose that (iv) holds. Noting that $T(\{\bar{x}\}, \bar{x}) = \{0\}$, by Proposition 3.1 we need only show that $L_f(\bar{x}) \cap Z \cap B(\bar{x}, \delta) = \{\bar{x}\}$ for some $\delta > 0$. Suppose to the contrary that there exists a sequence $\{x_k\}$ in $L_f(\bar{x}) \cap Z \setminus \{\bar{x}\}$ such that $x_k \to \bar{x}$. Without loss of generality we assume that $\frac{x_k - \bar{x}}{\|x_k - \bar{x}\|} \to h$ (passing to a subsequence if necessary). Thus, $h \in T(L_f(\bar{x}) \cap Z, \bar{x})$. It follows from Lemma 3.2 that

$$\langle f'(\bar{x}), h\rangle + \max_{y\in I_0(\bar{x})}[\langle\phi'_x(\bar{x}, y), h\rangle]_+ \le 0,$$

contradicting (iv) and $\|h\| = 1$. The proof is completed. □

4. Calmness for infinitely many smooth inequalities. Recently Henrion and Outrata [10] studied the calmness of infinitely many smooth inequalities. Let $C(Y)$ denote the Banach space of all continuous functions on Y equipped with the maximum norm, and consider the multifunction $M : C(Y) \rightrightarrows X$ defined by

$$M(g) := \{x \in X :\ \phi(x, y) \le -g(y) \quad \text{for all } y \in Y\} \quad \text{for all } g \in C(Y), \tag{4.1}$$

where X, Y, and $\phi(x, y)$ are as in section 3. For $\bar{g} \in C(Y)$ and $\bar{x} \in M(\bar{g})$, recall that M is calm at $(\bar{g}, \bar{x})$ if there exist $L,\ \delta \in (0,\ +\infty)$ such that

$$d(x, M(\bar{g})) \le L\|g - \bar{g}\| \quad \text{for all } g \in B(\bar{g}, \delta) \text{ and for all } x \in B(\bar{x}, \delta) \cap M(g).$$

We say that M is strongly calm at $(\bar{g}, \bar{x})$ if there exist $L,\ \delta \in (0,\ +\infty)$ such that

$$\|x - \bar{x}\| \le L\|g - \bar{g}\| \quad \text{for all } g \in B(\bar{g}, \delta) \text{ and for all } x \in B(\bar{x}, \delta) \cap M(g).$$

It is clear that M is strongly calm at $(\bar{g}, \bar{x})$ if and only if M is calm at $(\bar{g}, \bar{x})$ and $M(\bar{g}) \cap B(\bar{x}, \delta) = \{\bar{x}\}$ for some $\delta > 0$. Let $\Lambda := \{g \in C(Y) :\ g(y) \le 0 \quad \text{for all } y \in Y\}$ and $\bar{x} \in M(0)$. It is known (cf. [10]) that M is calm at $(0, \bar{x})$ if and only if there exist $L,\ \delta \in (0,\ +\infty)$ such that

$$d(x, M(0)) \le Ld(\phi(x, \cdot), \Lambda) \quad \text{for all } x \in B(\bar{x}, \delta).$$

Noting that

$$M(0) = Z \text{ and } d(\phi(x,\cdot),\Lambda) = \max_{y\in Y}[\phi(x,y)]_+,$$

it follows that M is calm at $(0,\bar{x})$ if and only if there exist $\eta,\ \delta \in (0,\ +\infty)$ such that

$$\eta d(x,Z) \le \max_{y\in Y}[\phi(x,y)]_+ \quad \text{for all } x \in B(\bar{x},\delta). \tag{4.2}$$

Setting $f(x) = 0$ for all $x \in X$ in (SIP), one sees that (4.2) means that $\bar{x}$ is a local weak sharp minimum of (SIP). Thus, by Theorems 3.1 and 3.2 and Proposition 3.1 we have the following characterizations for M to be calm at $(0,\bar{x})$.

THEOREM 4.1. *Let M be as in* (4.1) *and $\bar{x} \in M(0)$. Then the following statements are equivalent:*

(i) *M is calm at $(0,\bar{x})$.*

(ii) *There exist $\tau,\delta \in (0,\ +\infty)$ such that*

$$\hat{N}(M(0),u)\cap B_{X^*} \subset [0,\ \tau]\mathrm{co}\{\phi'_x(u,y) :\ y \in I_0(u)\} \quad \text{for all } u \in M(0)\cap B(\bar{x},\delta).$$

(iii) *There exist $\tau,\delta \in (0,\ +\infty)$ such that*

$$N(M(0),u)\cap B_{X^*} \subset [0,\ \tau]\mathrm{co}\{\phi'_x(u,y) :\ y \in I_0(u)\} \quad \text{for all } u \in M(0)\cap B(\bar{x},\delta).$$

(iv) *There exist $\tau,\delta \in (0,\ +\infty)$ such that*

$$N_c(M(0),u)\cap B_{X^*} \subset [0,\ \tau]\mathrm{co}\{\phi'_x(u,y) :\ y \in I_0(u)\} \quad \text{for all } u \in M(0)\cap B(\bar{x},\delta).$$

(v) *There exist $\tau,\delta \in (0,\ +\infty)$ such that*

$$d(h,T(M(0),u)) \le \tau \max_{y\in I_0(u)}[\langle\phi'_x(u,y),h\rangle]_+$$

for all $u \in M(0)\cap B(\bar{x},\delta)$ and $h \in X$.

(vi) *There exist $\tau,\delta \in (0,\ +\infty)$ such that*

$$d(h,T_c(M(0),u)) \le \tau \max_{y\in I_0(u)}[\langle\phi'_x(u,y),h\rangle]_+$$

for all $u \in M(0)\cap B(\bar{x},\delta)$ and $h \in X$.

(vii) *There exist $\tau,\delta \in (0,\ +\infty)$ such that*

$$\|x-u\| \le \tau \max_{y\in I_0(u)}[\langle\phi'_x(u,y),x-u\rangle]_+$$

for any $x \in B(\bar{x},\delta)$ and $u \in P_{M(0)}(x)$.

(viii) *$T(M(0),\bar{x}) = \{h \in X :\ \langle\phi'_x(\bar{x},y),h\rangle \le 0 \quad \text{for all } y \in I_0(\bar{x})\}$, and there exist $\tau,\delta \in (0,\ +\infty)$ such that for any $u \in M(0)\cap B(\bar{x},\delta)\setminus\{\bar{x}\}$*

$$\hat{N}(M(0),u)\cap B_{X^*} \subset [0,\ \tau]\mathrm{co}\{\phi'_x(u,y) :\ y \in I_0(u)\}.$$

In the remainder of this section, we assume that Y is a compact subset of R^n. Following Henrion and Outrata [10], let

$$\mathcal{J} := \{S \in \mathcal{K}(Y) :\ \exists x_i \overset{\mathrm{bd}M(0)\setminus\{\bar{x}\}}{\longrightarrow} \bar{x} \text{ such that } d_H(S,I_0(x_i)) \to 0\},$$

where $\mathcal{K}(Y)$ denotes the family of all compact subsets of Y and d_H denotes the Hausdorff distance between compact sets.

COROLLARY 4.1. *Let M be as in* (4.1) *and $\bar{x} \in M(0)$. Suppose that the following conditions are satisfied:*

1. $T(M(0),\bar{x}) = \{h \in X :\ \langle \phi'_x(\bar{x},y), h\rangle \le 0 \quad \textit{for all } y \in I_0(\bar{x})\}$.
2. *There exists* $\rho > 0$ *such that*

$$d(0, \mathrm{co}\{\phi'_x(\bar{x},y) :\ y \in S)\}) > \rho \quad \textit{for all } S \in \mathcal{J}.$$

Then M *is calm at* $(0,\bar{x})$.

Proof. We claim that there exists $\delta > 0$ such that

$$d(0, \mathrm{co}\{\phi'_x(x,y) :\ y \in I_0(x)\}) > \rho \quad \text{for all } x \in \mathrm{bd}(M(0)) \cap B(\bar{x},\delta) \setminus \{\bar{x}\}. \tag{4.3}$$

If this is not the case, then there exists a sequence $\{x_i\}$ in $\mathrm{bd}(M(0)) \setminus \{\bar{x}\}$ such that

$$x_i \to \bar{x} \ \text{ and } \ d(0, \mathrm{co}\{\phi'_x(x_i,y) :\ y \in I_0(x_i)\}) \le \rho \quad \text{for all } i \in \mathbb{N}.$$

Noting that X is of dimension m, it follows from the Caratheodory theorem that for each $i \in \mathbb{N}$ there exist $t_{ji} \ge 0$ and $y_{ji} \in I_0(x_i)$, $j = 1,\dots,m+1$, such that

$$\sum_{j=1}^{m+1} t_{ji} = 1 \ \text{ and } \ \left\| \sum_{j=1}^{m+1} t_{ji}\phi'_x(x_i,y_{ji}) \right\| \le \rho.$$

By (P2) and the compactness of Y, without loss of generality we assume that

$$t_{ji} \to t_j \ge 0 \ \text{ and } \ y_{ji} \to y_j \in I_0(\bar{x}),\ j = 1,\dots,m+1.$$

Hence,

$$\sum_{j=1}^{m+1} t_j = 1 \ \text{ and } \ \left\| \sum_{j=1}^{m+1} t_j\phi'_x(\bar{x},y_j) \right\| \le \rho. \tag{4.4}$$

Since the space of compact subsets of X endowed with the Hausdorff distance is itself compact, without loss of generality we assume that there exists $S_0 \in \mathcal{J}$ such that $d_H(I_0(x_i), S_0) \to 0$. It is clear that $y_j \in S_0$, $j = 1,\dots,m+1$. This and (4.4) imply that $d(0, \mathrm{co}\{\phi'_x(\bar{x},y) :\ y \in S_0)\}) \le \rho$, contradicting condition 2. Hence there exists $\delta > 0$ such that (4.3) holds. Recalling (cf. [26, Definition 7.25 and Theorem 10.31]) that $\Phi(x) = \max_{y\in Y} \phi(x,y)$ is regular and $\partial\Phi(x) = \mathrm{co}\{\phi'(x,y) :\ y \in I(x)\}$, it follows from (4.3) and [26, Proposition 10.3] that

$$N(M(0),x) = R_+\mathrm{co}\{\phi'_x(x,y) :\ y \in I_0(x)\} \quad \text{for all } x \in \mathrm{bd}(M(0)) \cap B(\bar{x},\delta) \setminus \{\bar{x}\}.$$

Let $x \in \mathrm{bd}(M(0)) \cap B(\bar{x},\delta) \setminus \{\bar{x}\}$ and $x^* \in N(M(0),x) \cap B_{X^*}$. Then there exist $t \in [0,\ +\infty)$ and $u^* \in \mathrm{co}\{\phi'_x(x,y) :\ y \in I_0(x)\}$ such that $x^* = tu^*$. This and (4.3) imply that $t < \frac{1}{\rho}$. Hence,

$$N(M(0),x) \cap B_{X^*} \subset \left[0,\ \frac{1}{\rho}\right] \mathrm{co}\{\phi'_x(x,y) :\ y \in I_0(x)\}.$$

It is clear that

$$N(M(0),x) \cap B_{X^*} = \{0\} \subset \left[0,\ \frac{1}{\rho}\right] \mathrm{co}\{\phi'_x(\bar{x},y) :\ y \in I_0(x)\} \quad \text{for all } x \in \mathrm{int}(M(0)).$$

Hence, (viii) of Theorem 4.1 holds, and so M is calm at $(0,\bar{x})$. The proof is completed. □

Remark 4.1. Corollary 4.1 is a slight improvement of [10, Theorem 4], which, in addition to all assumptions on Corollary 4.1, requires that $(x,y) \mapsto \phi(x,y)$ is continuously differentiable and $(x,y) \mapsto \phi'(x,y)$ is locally Lipschitz. Noting that

$$\mathcal{J} = \{S \subset Y : \exists x_i \overset{\mathrm{bd}M(0)\setminus\{\bar{x}\}}{\longrightarrow} \bar{x} \text{ such that } I_0(x_i) = S \text{ for all } i \in \mathbb{N}\}$$

when Y is a finite set, Corollary 4.1 recaptures [10, Theorem 3].

The following example shows that implication (viii)⇒(i) of Theorem 4.1 properly improves [10, Theorems 4 and 3]. Let $X = R^2$, $Y = \{0,1\}$, $\phi((s,t),0) = 0$, and $\phi((s,t),1) = s - t$ for all $(s,t) \in R^2$. Then $M(0) = \{(s,t) \in R^2 : \ s \le t\}$ and $I_0(x) = Y$ for any $x \in \mathrm{bd}(M(0))$. Let $\bar{x} = (0,0)$. Then $\mathcal{J} = \{Y\}$. Noting that

$$\mathrm{co}\{\phi'((s,t),y) : \ y \in Y\} = \{(u,-u) : \ 0 \le u \le 1\} \quad \text{for all } (s,t) \in R^2,$$

it follows that $d(0, \mathrm{co}\{\phi'_x(\bar{x},y) : \ y \in S\}) = 0$ for all $S \in \mathcal{J}$. Thus, Corollary 4.1 and so [10, Theorems 4 and 3] are not applicable. On the other hand, noting that $\mathrm{bd}(M(0)) = \{(s,s) : \ s \in R\}$,

$$T(M(0),(s,s)) = M(0) \ \text{ and } \ N(M(0),(s,s)) = \{(t,-t) : \ t \ge 0\} \quad \text{for all } s \in R,$$

one can see that (viii) of Theorem 4.1 holds. Hence, applying implication (viii)⇒(i) of Theorem 4.1, one obtains that M is calm at $(0,\bar{x})$.

We conclude with characterizations for M to be strongly calm at $(0,\bar{x})$.

THEOREM 4.2. *Let M be as in* (4.1) *and $\bar{x} \in M(0)$. Then the following statements are equivalent:*

(i) *M is strongly calm at $(0,\bar{x})$.*

(ii) *There exists $\tau \in (0, +\infty)$ such that*

$$B_{X^*} \subset [0, \tau]\mathrm{co}\{\phi'_x(\bar{x},y) : \ y \in I_0(\bar{x})\}.$$

(iii) $X^* = R_+\mathrm{co}\{\phi'_x(\bar{x},y) : \ y \in I_0(\bar{x})\}$.

(iv) *There exists $\tau \in (0, +\infty)$ such that*

$$\|h\| \le \tau \max_{y \in I_0(\bar{x})} [\langle \phi'_x(\bar{x},y), h\rangle]_+ \quad \text{for all } h \in X.$$

(v) $\{h \in X : \ \langle \phi'_x(\bar{x},y), h\rangle \le 0 \quad \textit{for all } y \in I_0(\bar{x})\} = \{0\}$.

Proof. Noting that M is strongly calm at $(0,\bar{x})$ if and only if $\bar{x}$ is a local sharp minimum of (SIP) with $f \equiv 0$, (i)⇔(ii)⇔(iv)⇔(v) are immediate from Theorem 3.3. It is clear that (ii)⇒(iii)⇒(v) hold. The proof is completed. ∎

REFERENCES

[1] B. BROSOWSKI, *Parametric Semi-Infinite Optimization*, Verlag Peter Lang, Frankfurt, 1982.

[2] J. V. BURKE AND S. DENG, *Weak sharp minima revisited Part* I: *Basic theory*, Control Cybern., 31 (2002), pp. 439–469.

[3] J. V. BURKE AND M. C. FERRIS, *Weak sharp minima in mathematical programming*, SIAM J. Control Optim., 31 (1993), pp. 1340–1359.

[4] L. CROMME, *Strong uniqueness, a far-reaching criterion for the convergence analysis of iterative procedures*, Numer. Math., 29 (1978), pp. 179–193.

[5] M. C. FERRIS, *Weak Sharp Minima and Penalty Functions in Mathematical Programming*, Ph.D. thesis, University of Cambridge, Cambridge, 1988.

[6] A. GOBERNA AND M. A. LOPEZ, *Linear Semi-Infinite Optimization*, John Wiley & Sons, Chichester, 1998.

[7] A. Goberna and M. A. Lopez, eds., *Semi-infinite Programming—Recent Advances*, Kluwer, Boston, 2001.
[8] R. Henrion and A. Jourani, *Subdifferential conditions for calmness of convex constraints*, SIAM J. Optim., 13 (2002), pp. 520–534.
[9] R. Henrion, A. Jourani, and J. Outrata, *On the calmness of a class of multifunctions*, SIAM J. Optim., 13 (2002), pp. 603–618.
[10] R. Henrion and J. Outrata, *Calmness of constraint systems with applications*, Math. Program., 104 (2005), pp. 437–464.
[11] R. Hettich and K. O. Kortanek, *Semi-infinite programming: Theory, methods, and applications*, SIAM Rev., 35 (1993), pp. 380–429.
[12] K. Jittorntrum and M. R. Osborne, *Strong uniqueness and second order convergence in nonlinear discrete approximation*, Numer. Math., 34 (1980), pp. 439–455.
[13] H. Th. Jongen, J.-J. Ruuckmann, and O. Stein, *Generalized semi-infinite optimization: A first order optimality condition and examples*, Math. Program., 83 (1998), pp. 145–158.
[14] D. Klatte and R. Henrion, *Regularity and stability in nonlinear semi-infinite optimization, Semi-infinite programming*, Nonconvex Optim. Appl., 25 (1998), pp. 69–102.
[15] D. Klatte and B. Kummer, *Nonsmooth Equations in Optimization*, in Regularity, Calculus, Methods and Applications, Nonconvex Optim. Appl. 60, Kluwer Academic, Dordrecht, 2002.
[16] A. Lewis and J. S. Pang, *Error bounds for convex inequality systems*, in Proceedings of the Fifth Symposium on Generalized Convexity, Luminy, 1996, J.-P. Crouzeix, J.-E. Martinez-Legaz, and M. Volle, eds., Kluwer Academic, Dordrecht, 1997, pp. 75–110.
[17] W. Li, *Sharp Lipschitz constants for basic optimal solutions and basic feasible solutions of linear programs*, SIAM J. Control Optim., 32 (1994), pp. 140–153.
[18] B. S. Mordukhovich, *Variational Analysis and Generalized Differentiation* I, Basic Theory, Springer-Verlag, Berlin, 2006.
[19] B. S. Mordukhovich, *Variational Analysis and Generalized Differentiation* II, Applications, Springer-Verlag, Berlin, 2006.
[20] G. Nurnberger, *Global unicity in semi-infinite optimization*, Numer. Funct. Anal. Optim., 8 (1985), pp. 173–191.
[21] E. Polak, *Optimization*, Springer-Verlag, New York, 1997.
[22] M. R. Osborne and R. S. Womersley, *Strong uniqueness in sequential linear programming*, J. Aust. Math. Soc. Ser. B, 31 (1990), pp. 379–384.
[23] J. S. Pang, *Error bounds in mathematical programming*, Math. Program. Ser. B, 79 (1997), pp. 299–332.
[24] B. T. Polyak, *Introduction to Optimization*, Optimization Software, New York, 1987.
[25] R. Reemtsen and J.-J. Rckmann, eds., *Semi-Infinite Programming*, Kluwer Academic, Boston, 1998.
[26] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
[27] O. Stein, *Bi-level Strategies in Semi-infinite Programming*, Kluwer, Boston, 2003.
[28] M. Studniarski and D. E. Ward, *Weak sharp minima: Characterizations and sufficient conditions*, SIAM J. Control Optim., 38 (1999), pp. 219–236.
[29] F. G. Vázquez and J.-J. Rückmann, *Extensions of the Kuhn-Tucker constraint qualification to generalized semi-infinite programming*, SIAM J. Optim., 15 (2005), pp. 926–937.
[30] R. S. Womersley, *Local properties of algorithms for minimizing nonsmooth composite functions*, Math. Program., 32 (1985), pp. 69–89.
[31] C. Zalinescu, *Weak sharp minima, well-behaving functions and global error bounds for convex inequalities in Banach spaces*, in Proceedings of the 12th Baikal International Conference on Optimization Methods and Their Application, Institute of System Dynamics and Control Theory of SB RAS, Irkutsk, Russia, 2001, pp. 272–284.
[32] C. Zualinescu, *Sharp estimates for Hoffman's constant for systems of linear inequalities and equalities*, SIAM J. Optim., 14 (2003), pp. 517–533.
[33] X. Y. Zheng and K. F. Ng, *Metric regularity and constraint qualifications for convex inequalities on Banach spaces*, SIAM J. Optim., 14 (2003), pp. 757–772.
[34] X. Y. Zheng and X. Q. Yang, *Lagrange multipliers in nonsmooth semi-infinite optimization problems*, Math. Oper. Res., 32 (2007), pp. 168–181.

SIAM J. OPTIM.
Vol. 18, No. 2, pp. 589–612

SECOND-ORDER NECESSARY CONDITIONS FOR NONLINEAR OPTIMIZATION PROBLEMS WITH ABSTRACT CONSTRAINTS: THE DEGENERATE CASE*

HELMUT GFRERER†

Abstract. In this paper we derive second-order necessary conditions for optimality for an optimization problem with abstract constraints in Banach spaces. Results for the nondegenerate case derived earlier [H. Gfrerer, *SIAM J. Control Optim.*, 45 (2006), pp. 972–997] are extended to the degenerate case. For the mathematical programming problem, where the constraints are given by equality and finitely many inequality constraints, our approach applies to the degenerate case, when the equality constraints are not regular; our results appear to be new even in this special case. Our second-order necessary conditions are contained in the gap between the standard necessary and sufficient conditions, where the only difference is the change from a nonstrict to a strict inequality. Our results are formulated in such a way to be applicable also to vector optimization problems.

Key words. second-order optimality conditions, abstract constraints, vector optimization

AMS subject classifications. 49K27, 90C29, 90C30

DOI. 10.1137/050641387

1. Introduction.

In this paper we study necessary second-order optimality conditions for minimization problems with abstract constraints of the form

$$g(x) \in K, \tag{1.1}$$

where the constraint mapping $g : X \to V$ carries a Banach space X into another Banach space V, and K is a closed convex subset of V.

The feasible sets of various optimization problems can be formulated in the form (1.1) in a natural way. For instance, the possibly infinite-dimensional mathematical programming problem with finitely many inequality constraints $g_i(x) \le 0$, $i = 1, \dots, m$, and an equality constraint $G(x) = 0$, where $G : X \to \hat{V}$ maps X into another Banach space $\hat{V}$, fits into the scheme with $V = \mathbb{R}^m \times \hat{V}$, $K = \mathbb{R}^m_- \times \{0\}$, and $g = (g_1, \dots, g_m, G)$. Other examples are provided by semi-infinite programming problems, semidefinite programming problems, and optimal control problems. We refer to the monograph [10] for a comprehensive overview.

When studying necessary optimality conditions at a point $\bar{x}$ satisfying (1.1), usually a condition on the constraints is needed, since otherwise the necessary conditions trivially hold and therefore their utility for describing optimality is very limited. One classical condition in this setting is Robinson's condition [30]:

$$0 \in \operatorname{int}\left(g(\bar{x}) + g'(\bar{x})X - K\right). \tag{1.2}$$

Note that in the case of the finite-dimensional mathematical programming problem, condition (1.2) reduces to the classical Mangasarian–Fromovitz constraint qualification. Under Robinson's condition, the structure of the set of tangent directions of the

*Received by the editors September 28, 2005; accepted for publication (in revised form) January 18, 2007; published electronically July 4, 2007. This work was partially supported by the Austrian Science Fund (FWF) under grant SFB F013/F1309.

http://www.siam.org/journals/siopt/18-2/64138.html

†Institute of Computational Mathematics, Johannes Kepler University Linz, A-4040 Linz, Austria (gfrerer@numa.uni-linz.ac.at).

constraints is well established (see, e.g., [10]) and second-order conditions have been formulated by several authors; see [9], [14], [17], [23], [28], [29].

In a recent paper [15], we presented second-order necessary optimality conditions, which are shown to be the best possible in a certain sense, under the nondegeneracy assumption

$$\operatorname{int}(g'(\bar{x})X - K) \neq \emptyset. \tag{1.3}$$

This condition is clearly weaker than Robinson's condition (1.2). For instance, for the mathematical programming problem, condition (1.3) reduces to a condition on the equality constraint only, namely, the well-known Lyusternik condition $G'(\bar{x})X = \hat{V}$, which has been used by several authors (see, e.g., [7], [8], [27], [16]). However, it is well known that there exist second-order necessary conditions for the mathematical programming problem which do not require the Lyusternik condition $G'(\bar{x})X = \hat{V}$ to be satisfied. A generalization of the classical Lyusternik theorem for nonregular mappings $G'(\bar{x})$ was first derived and proved in [12]. Tret'yakov [33] and Avakov [4], [5] presented optimality conditions for this nonregular case. These results have been extended by several authors; see, e.g., [1], [2], [6], [11], [19], [25], [26].

In this paper we will derive second-order necessary optimality conditions in the degenerate case, i.e., condition (1.3) is dropped. The obtained results appear to be new even in the special case of the mathematical programming problem. Our approach is essentially based on the ideas presented in [15]. We will use both an observation made by Robinson [32]—that a certain multifunction built by the objective and the constraints obtains a singular behavior at a local minimizer—and an accurate characterization of metric regularity of this multifunction by means of a certain signed distance function.

With this approach, when dealing with general constraints of the form (1.1), it does not require great effort to consider also the case of general objective functions. Thus the problem we consider in this paper is given by

$$\text{L-minimize } f(x) \quad \text{subject to} \quad g(x) \in K, \tag{P}$$

where $f : X \to U$ is a mapping from the Banach space X to another Banach space U and where $L \subset U$ is a closed convex cone with nonempty interior, $\operatorname{int} L \neq \emptyset$. We define different kinds of local L-minimizers as follows.

Definition 1.1. *An element $\bar{x} \in X$ is called a* local weak minimizer *for* (P) *if $g(\bar{x}) \in K$ and if there exists a neighborhood N of $\bar{x}$ such that for each $x \in N$ with $g(x) \in K$, one has $f(x) - f(\bar{x}) \notin -\operatorname{int} L$. A local weak minimizer $\bar{x}$ is called a* strict local minimizer *for* (P) *if for each $x \in N \setminus \{\bar{x}\}$ with $g(x) \in K$, one has $f(x) - f(\bar{x}) \notin -L$. Finally, a weak local minimizer $\bar{x}$ is called an* essential local minimizer of second order *for problem* (P) *if there exists some real $\beta > 0$ such that*

$$\max\{d(f(x) - f(\bar{x}), -L), d(g(x), K)\} \geq \beta \|x - \bar{x}\|^2 \quad \forall x \in N. \tag{1.4}$$

Of course, each essential local minimizer of second order is also a strict local minimizer.

Note that (P) includes the very common problem of constrained scalar minimization, for which $U = \mathbb{R}$ and $L = \mathbb{R}_+$. Local weak minimizers for (P) then amount to usual local minimizers, and for essential local minimizers the so-called *quadratic growth condition* is satisfied:

$$f(x) - f(\bar{x}) \geq \beta \|x - \bar{x}\|^2 \quad \forall x \in N : \ g(x) \in K.$$

Another important particular case, for instance, is the Pareto maximization optimization problem with $U = \mathbb{R}^p$, $L = \mathbb{R}^p_-$.

For second-order optimality conditions for problem (P) in the multicriteria case we refer to [8], [13], [15], [20], [21], [22].

Given a feasible point $\bar{x} \in g^{-1}(K)$, fixed throughout this paper, we will now define a certain multifunction associated with (P) and $\bar{x}$. Let $h : X \to U \times V$ be defined by

$$h(x) := (f(x) - f(\bar{x}),\ g(x)), \qquad C := (-L) \times K, \qquad Y := U \times V. \tag{1.5}$$

Then the multifunction $\Gamma : X \rightrightarrows Y$, given by

$$\Gamma(x) := h(x) - C, \tag{1.6}$$

will form the basis of our investigations. Throughout this paper we will use the following smoothness assumption on h.

ASSUMPTION 1. *h is Fréchet differentiable at $\bar{x}$ and for some radius $\bar{r} > 0$ and some scalar $\eta \geq 0$ we have*

$$\|h(x_1) - h(x_2) - h'(\bar{x})(x_1 - x_2)\| \leq \eta \max\{\|x_1 - \bar{x}\|, \|x_2 - \bar{x}\|\}\|x_1 - x_2\|$$

for all $x_1, x_2 \in \bar{x} + \bar{r}\mathcal{B}_X$.

In what follows we will use the norm $\|(u, v)\| := \max\{\|u\|, \|v\|\}$ on the product space $Y = U \times V$.

Our notation is fairly standard. In a normed space Z, $\mathcal{B}_Z := \{z \in Z : \|z\| \leq 1\}$ denotes the closed unit ball and $\mathcal{S}_Z := \{z \in Z : \|z\| = 1\}$ denotes the unit sphere. The topological dual space is denoted by Z^*. $\langle z^*, z\rangle$ is the value $z^*(z)$ of the linear functional $z^* \in Z^*$ at $z \in Z$. For a set $D \subset Z$ we denote by $\sigma_D(\cdot)$ its support function, i.e., $\sigma_D(z^*) := \sup_{z\in D}\langle z^*, z\rangle$, and by $d(\cdot, D)$ the distance function, i.e., $d(z, D) = \inf_{y\in D} \|z - y\|$. For a convex set $D \subset Z$ we denote by $T_D(z)$ (respectively, $N_D(z)$) the common tangent cone (respectively, normal cone) of convex analysis at a point $z \in D$, i.e., we have $N_D(z) := \{z^* \in Z^* : \langle z^*, \zeta - z\rangle \leq 0 \text{ for all } \zeta \in D\}$ and

$$T_D(x) = \left\{ s : \liminf_{t\to 0_+} \frac{d(z + ts, D)}{t} = 0 \right\} = \left\{ s : \limsup_{t\to 0_+} \frac{d(z + ts, D)}{t} = 0 \right\}.$$

If W is another normed space, we denote by $L(Z, W)$ the space of all continuous linear operators from Z into W. If $A \in L(Z, W)$, then $A^* : W^* \to Z^*$ denotes the adjoint operator of A. Finally, we denote by $\mathbb{T}$ the set of all sequences $(t_n) \to 0_+$.

Fritz–John-type optimality conditions for problem (P) can be written in the form

$$f'(\bar{x})^*u^* + g'(\bar{x})^*v^* = 0, \quad 0 \neq (u^*, v^*) \in L^* \times N_K(g(\bar{x})) \subset U^* \times V^*, \tag{1.7}$$

where $L^* := \{u^* \in U^* : \langle u^*, u\rangle \geq 0 \text{ for all } u \in L\}$ is the dual cone of the cone L. Setting $y^* := (u^*, v^*)$ and using the notation of h and C, the condition (1.7) can also be written more shortly as

$$h'(\bar{x})^*y^* = 0, \quad y^* \in N_C(h(\bar{x})), \quad y^* \neq 0. \tag{1.8}$$

In what follows we will denote the set of multipliers y^* satisfying the Fritz–John conditions (1.8) by Λ_{FJ}. It should be noted that in general Λ_{FJ} may be empty. An additional condition has to be imposed to ensure the existence of a nontrivial multiplier y^* at a local weak minimizer for (P).

2. Preliminaries. In this section we will recapitulate partially the basic theory on second-order optimality conditions as presented in [15].

In a very general form, these conditions can be formulated by means of a function $\hat{d}_C(y, A, \kappa) : Y \times L(X, Y) \times \mathbb{R} \to \mathbb{R}$ given by

$$\hat{d}_C(y, A, \kappa) := \sup_{y^* \in \mathcal{S}_{Y^*}} \{\langle y^*, y\rangle - \sigma_C(y^*) - \kappa \|A^* y^*\|\}.$$

THEOREM 2.1 (see [15, Theorem 3.2]). *Suppose that Assumption* 1 *is satisfied at* $\bar{x}$. *If* $\bar{x}$ *is a local weak minimizer for* (P), *then*

$$\liminf_{\substack{x \to \bar{x} \\ \tau \to 0_+}} \frac{\hat{d}_C(h(x), h'(\bar{x}), \tau\|x - \bar{x}\|)}{\|x - \bar{x}\|^2} \geq 0. \tag{2.1}$$

Moreover, a feasible point $\bar{x}$ *is an essential local minimizer of second order for* (P) *if and only if*

$$\liminf_{\substack{x \to \bar{x} \\ \tau \to 0_+}} \frac{\hat{d}_C(h(x), h'(\bar{x}), \tau\|x - \bar{x}\|)}{\|x - \bar{x}\|^2} > 0. \tag{2.2}$$

The following theorem gives further details on the optimality conditions of Theorem 2.1 (see [15, Theorems 3.5, 3.6]).

THEOREM 2.2.

1. *Suppose that a feasible point* $\bar{x}$ *is not an essential local minimizer of second order for* (P). *Then there exists a twice continuously differentiable mapping* $\delta h := (\delta f, \delta g)$ *satisfying* $\delta h(\bar{x}) = 0$, $\delta h'(\bar{x}) = 0$, *and* $\delta h''(\bar{x}) = 0$, *such that* $\bar{x}$ *is not a local weak minimizer for* (P) *with* f *and* g *replaced by* $f + \delta f$ *and* $g + \delta g$, *respectively.*

2. *Assume that at a feasible point* $\bar{x}$ *condition* (2.1) *holds, and assume that*

$$\operatorname{int}(h'(\bar{x})X - C) \neq \emptyset. \tag{2.3}$$

Then there exists a mapping $\delta h = (\delta f, \delta g) : X \to Y$ *with* $\delta h(x) = \psi(\|x - \bar{x}\|)y$, *where* $y \in Y$ *and* $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ *is a twice continuously differentiable function satisfying* $\psi(0) = \psi'(0) = \psi''(0) = 0$, *such that* $\bar{x}$ *is a strict local minimizer for* (P) *with* f *and* g *replaced by* $f + \delta f$ *and* $g + \delta g$, *respectively.*

Remark. It follows from the proof of [15, Theorem 3.6] that the assertion of the second part of Theorem 2.2 does not depend on the special form of h and C, respectively. Indeed, for any closed convex set $C \subset Y$ and any mapping $h : X \to Y$, differentiable at some point $\bar{x}$ satisfying $h(\bar{x}) \in C$ and conditions (2.1) and (2.3), there exist a neighborhood N of $\bar{x}$ and a mapping $\delta h(x) = \psi(\|x - \bar{x}\|)y$ of the same kind as in the second part of Theorem 2.2, such that $d\big((h + \delta h)(x), C\big) > 0$ for all $x \in N \setminus \{\bar{x}\}$.

DEFINITION 2.3. *We call* $\bar{x}$ nondegenerate *for the problem* (P) *if* $\operatorname{int}(h'(\bar{x})X - C) \neq \emptyset$. *Conversely, if* $\operatorname{int}(h'(\bar{x})X - C) = \emptyset$, *the element* $\bar{x}$ *is said to be* degenerate *for* (P).

The following theorem states some geometrical properties of the function $\hat{d}_C$: It can be treated as a signed distance function for certain sets.

THEOREM 2.4 (see [15, Proposition 2.6]). *For each* $A \in L(X, Y)$, *each* $y \in Y$, *and each* $\kappa \geq 0$ *let* $D_C(y, A, \kappa)$ *be given by* $D_C(y, A, \kappa) := y + \kappa A\mathcal{B}_X - C$. *Then one*

has

$$\hat{d}_C(y,A,\kappa)=\begin{cases} d(0,D_C(y,A,\kappa)) & \text{if } 0\notin \operatorname{cl} D_C(y,A,\kappa),\\ 0 & \text{if } 0\in \operatorname{bd} D_C(y,A,\kappa),\\ -\sup\{\rho\,:\,\rho\mathcal{B}_Y\subset D_C(y,A,\kappa)\} & \text{if } 0\in \operatorname{int} D_C(y,A,\kappa).\end{cases}$$

It follows easily from the definition that $\hat{d}_C(\cdot,A,\kappa)$ is Lipschitz continuous with constant 1. Moreover we have the following property.

LEMMA 2.5 (see [15, Lemma 2.8]). *Let $A\in L(X,Y)$, $y\in Y$, and $\kappa\geq 0$ be such that $\hat{d}_C(y,A,\kappa)<0$. Then*

$$d(0,A^{-1}(C-y))\leq \frac{\kappa}{d(y,C)-\hat{d}_C(y,A,\kappa)}d(y,C).$$

For the sake of completeness we mention some material from [15] also used in this paper. First, let us recall the notion of (local) metric regularity.

DEFINITION 2.6. *Let $\Psi: X\rightrightarrows Y$ be a set-valued map, $\bar{y}\in\Psi(\bar{x})$. The multifunction Ψ is called* metrically regular near $(\bar{x},\bar{y})$ *if there are neighborhoods $N_{\bar{x}}$, $N_{\bar{y}}$ of $\bar{x}$, $\bar{y}$, respectively, and some $k>0$ such that $d(x,\Psi^{-1}(y))\leq k\,d(y,\Psi(x))$ for each $(x,y)\in N_{\bar{x}}\times N_{\bar{y}}$.*

The following two theorems are the base for the necessary optimality conditions of Theorem 2.1 and are also of substantial significance for this paper.

THEOREM 2.7 (see [15, Proposition 2.4]). *Let (x_n) be a sequence converging to $\bar{x}$ such that for each n, Γ is metrically regular near $(x_n,0)$. Then $\bar{x}$ is not a local weak minimizer for* (P).

THEOREM 2.8 (see [15, Proposition 2.10]). *Let $\hat{x}\in X$ be given and suppose that there exist a continuous linear mapping $A\in L(X,Y)$, a vector $x^0\in X$, and scalars $R>0$, $\hat{\kappa}\geq 0$, and $\gamma>0$, such that the following conditions are satisfied:*

$$\|h(x')-h(x)-A(x'-x)\|\leq\gamma\|x'-x\|\ \forall x,x'\in\hat{x}+R\mathcal{B}_X, \tag{2.4}$$

$$h(\hat{x})+A(x^0-\hat{x})\in C,\ r:=\|x^0-\hat{x}\|<R/2, \tag{2.5}$$

$$2\gamma(\hat{\kappa}+3r)+\hat{d}_C(h(\hat{x}),A,\hat{\kappa})<0. \tag{2.6}$$

Then there exists some $\tilde{x}\in x^0+r\mathcal{B}_X$, such that $h(\tilde{x})\in C$ and Γ is metrically regular near $(\tilde{x},0)$. Moreover, $\hat{d}_C(h(\tilde{x}),A,\hat{\kappa}+\|\tilde{x}-\hat{x}\|)\leq\hat{d}_C(h(\tilde{x}),A,\hat{\kappa})+\gamma\|\tilde{x}-\hat{x}\|$.

3. A general necessary condition. The following theorem states an abstract necessary optimality condition for problem (P). In some sense it is contained in the gap between the necessary and sufficient optimality conditions of Theorem 2.1.

THEOREM 3.1. *Assume that Assumption* 1 *holds. Further assume that $\bar{x}$ is a local weak minimizer, but not an essential local minimizer of second order, and let $(z_n)\subset\mathcal{S}_X$, $(t_n)\in\mathbb{T}$, and $(\tau_n)\in\mathbb{T}$ be sequences such that*

$$\limsup_{n\to\infty} t_n^{-2}\hat{d}_C(h(\bar{x}+t_nz_n),h'(\bar{x}),\tau_nt_n)\leq 0. \tag{3.1}$$

Further assume that there is a sequence $(A_n)\subset L(X,Y)$ of continuous linear operators mapping X into Y such that, together with some positive scalars $\gamma',R'>0$ and a sequence $(\varphi_n')\in\mathbb{T}$, one has

$$\begin{aligned}&\|h(x')-h(x)-A_n(x'-x)\|\\ &\qquad\leq\big(\varphi_n't_n+\gamma'\max\{\|\bar{x}+t_nz_n-x'\|,\|\bar{x}+t_nz_n-x\|\}\big)\|x'-x\|\end{aligned} \tag{3.2}$$

for all $x', x \in \bar{x} + t_n(z_n + R'\mathcal{B}_X)$ *and for each* n. *Then for each* $T > 0$ *one has*

$$\liminf_{n\to\infty} t_n^{-2}\hat{d}_C(h(\bar{x}+t_nz_n), A_n, Tt_n) \geq 0. \tag{3.3}$$

Proof. By Theorem 2.4, for each n we can find some element $\delta z_n \in \mathcal{B}_X$ such that

$$\begin{aligned}\delta_n :&= t_n^{-2} d(h(\bar{x}+t_nz_n) + t_n\tau_n h'(\bar{x})\delta z_n, C)\\ &\leq t_n^{-2}\max\{\hat{d}_C(h(\bar{x}+t_nz_n), h'(\bar{x}), \tau_n t_n), 0\} + \frac{1}{n},\end{aligned}$$

and from (3.1) it follows that $\delta_n \to 0$. Let $z_n' := z_n + \tau_n\delta z_n$. Then $\|z_n'\| \leq 1+\tau_n$, and using Assumption 1 we obtain

$$\begin{aligned}\delta_n' &:= t_n^{-2} d(h(\bar{x}+t_nz_n'), C)\\ &\leq t_n^{-2}\big(d(h(\bar{x}+t_nz_n) + t_n\tau_n h'(\bar{x})\delta z_n, C)\\ &\qquad + \|h(\bar{x}+t_nz_n') - h(\bar{x}+t_nz_n) - t_n\tau_n h'(\bar{x})\delta z_n\|\big)\\ &\leq \delta_n + t_n^{-2}\eta\max\{\|t_nz_n'\|, \|t_nz_n\|\}\|t_n(z_n - z_n')\| \leq \delta_n + \eta(1+\tau_n)\tau_n,\end{aligned}$$

yielding $\delta_n' \to 0$. We will now prove by contraposition that (3.3) holds for arbitrarily fixed $T > 0$. Assume on the contrary that

$$\liminf_{n\to\infty} t_n^{-2}\hat{d}_C(h(\bar{x}+t_nz_n), A_n, Tt_n) \leq -2\epsilon < 0$$

for some $T > 0$. Using the Lipschitz continuity of $\hat{d}_C(\cdot, A_n, Tt_n)$, Theorem 2.4, and condition (3.2) we obtain

$$\begin{aligned}\hat{d}_C(h(\bar{x}+t_nz_n'), A_n, Tt_n) &\leq \hat{d}_C(h(\bar{x}+t_nz_n) + \tau_nt_nA_n\delta z_n, A_n, Tt_n)\\ &\qquad + \|h(\bar{x}+t_nz_n') - h(\bar{x}+t_nz_n) - \tau_nt_nA_n\delta z_n\|\\ &\leq \hat{d}_C(h(\bar{x}+t_nz_n), A_n, (T+\tau_n)t_n)\\ &\qquad + (\varphi_n't_n + \gamma'\tau_nt_n\|\delta z_n\|)\tau_nt_n\|\delta z_n\|\\ &= \hat{d}_C(h(\bar{x}+t_nz_n), A_n, (T+\tau_n)t_n) + o(t_n^2).\end{aligned}$$

Next define $T_n := T + \tau_n$ and $\mu_n := \max\{\tau_n, \varphi_n', \frac{T_n\delta_n'}{\epsilon}\}$ for each n. Since $\mu_n \to 0$, by passing to a subsequence if necessary, we may assume that

$$\begin{aligned}t_n^{-2}\hat{d}_C(h(\bar{x}+t_nz_n'), A_n, T_nt_n) &\leq -\epsilon,\\ (2+8\gamma')(\mu_nT_n + 3\mu_n^2) - \epsilon &< 0,\\ \tau_n + 3\mu_n &\leq R'\end{aligned}$$

for each n. Now let n be arbitrarily fixed. We will now show that the assumptions of Theorem 2.4 hold with data $\hat{x} = \bar{x} + t_nz_n'$, $A = A_n$, $R = 3\mu_nt_n$, $\hat{\kappa} = T_nt_n$, and $\gamma = (1+4\gamma')\mu_nt_n$. Since $\|t_n(z_n' - z_n)\| + R \leq t_n(\tau_n + 3\mu_n) \leq t_nR'$ it follows from (3.2) for all $x, x' \in \bar{x} + t_nz_n' + R\mathcal{B}_X \subset \bar{x} + t_n(z_n + R'\mathcal{B}_X)$ that

$$\begin{aligned}&\|h(x') - h(x) - A_n(x'-x)\|\\ &\qquad \leq (t_n\varphi_n' + \gamma'\max\{\|\bar{x}+t_nz_n - x'\|, \|\bar{x}+t_nz_n - x\|\})\|x - x'\|\\ &\qquad \leq (t_n\varphi_n' + \gamma'(R + t_n\|z_n - z_n'\|))\|x - x'\|\\ &\qquad \leq (\varphi_n' + \gamma'(3\mu_n + \tau_n))t_n\|x - x'\| \leq (1+4\gamma')\mu_nt_n\|x - x'\|,\end{aligned}$$

and hence (2.4) holds. Using Lemma 2.5 we have

$$d(0, A_n^{-1}(h(\bar{x}+t_nz_n')-C)) \le \frac{T_nt_nd(h(\bar{x}+t_nz_n'),C)}{d(h(\bar{x}+t_nz_n'),C)-\hat{d}_C(h(\bar{x}+t_nz_n'),A_n,T_nt_n)} \le \frac{T_nt_n\delta_n't_n^2}{\delta_n't_n^2+\epsilon{t_n}^2} < \frac{T_n\delta_n'}{\epsilon}t_n \le \mu_nt_n.$$

Thus there exists some x^0 such that

$$h(\bar{x}+t_nz_n')+A_n(x^0-(\bar{x}+t_nz_n')) \in C, \quad r := \|x^0-(\bar{x}+t_nz_n')\| \le \mu_nt_n = \frac{R}{3},$$

showing the validity of (2.5). Finally, we have

$$2\gamma(\hat{\kappa}+3r)+\hat{d}_C(h(\bar{x}+t_nz_n'),A_n,\hat{\kappa}) \le 2(1+4\gamma')\mu_nt_n(T_nt_n+3\mu_nt_n)-\epsilon t_n^2 = \big((2+8\gamma')(\mu_nT_n+3\mu_n^2)-\epsilon\big)t_n^2 < 0,$$

yielding (2.6). Thus we can apply Theorem 2.4 to establish the existence of some $\tilde{x}_n \in \bar{x}+t_nz_n'+3\mu_nt_n\mathcal{B}_X$ such that $h(\tilde{x}_n) \in C$ and the multifunction $h(\cdot)-C$ is metrically regular near $(\bar{x},0)$. This holds for each n, and since $\tilde{x}_n \to \bar{x}$ we conclude from Theorem 2.7 that $\bar{x}$ is not a local weak minimizer, a contradiction. □

It is easy to show that if h is continuously differentiable in a neighborhood of $\bar{x}$, then for a sequence $(A_n) \subset L(X,Y)$ of linear operators satisfying condition (3.2) one has $\|h'(\bar{x}+t_nz_n)-A_n\| \le \varphi_n't_n = o(t_n)$. The converse is also true if h is sufficiently smooth near $\bar{x}$.

LEMMA 3.2. *Suppose that h is continuously differentiable in some ball $\bar{x}+\rho\mathcal{B}_X$ around $\bar{x}$. Further suppose that either h is twice Fréchet differentiable at $\bar{x}$ or that $h'(\cdot)$ is Lipschitz continuous in $\bar{x}+\rho\mathcal{B}_X$. Then, for any sequences $(z_n) \subset \mathcal{S}_X$, $(t_n) \in \mathbb{T}$, and $(A_n) \subset L(X,Y)$ such that $\|h'(\bar{x}+t_nz_n)-A_n\| = o(t_n)$, there exist a sequence $(\varphi_n') \in \mathbb{T}$ and positive reals γ' and R' such that condition* (3.2) *holds for all n sufficiently large.*

Proof. Let $R' > 0$ be arbitrarily chosen and consider n chosen so large that $t_n(z_n+R'\mathcal{B}_X) \subset \rho\mathcal{B}_X$. Consider an arbitrary linear functional $y^* \in \mathcal{B}_{Y^*}$. For every pair $x, x' \in \bar{x}+t_n(z_n+R'\mathcal{B}_X)$, by the mean-value theorem, there exists some element ξ belonging to the line segment $[x,x']$ such that $\langle y^*, h(x')-h(x)\rangle = \langle y^*, h'(\xi)(x'-x)\rangle$ and

$$\langle y^*, h(x')-h(\bar{x})-A_n(x'-x)\rangle = \langle y^*, (h'(\xi)-A_n)(x'-x)\rangle \le \|y^*\|\|h'(\xi)-A_n\|\|x'-x\|$$

follows. Now, in order to prove the lemma it is sufficient to show the bound

$$\|h'(\xi)-A_n\| \le \varphi_n't_n+\gamma'\max\{\|\bar{x}+t_nz_n-x'\|, \|\bar{x}+t_nz_n-x\|\} \tag{3.4}$$

for some constant γ' and some sequence $(\varphi_n') \in \mathbb{T}$. When h is twice Fréchet differentiable at $\bar{x}$ we have

$$\|h'(\xi)-h'(\bar{x}+t_nz_n)\| \le \|h'(\xi)-(h'(\bar{x})+h''(\bar{x})(\xi-\bar{x}))\| + \|h''(\bar{x})(\xi-(\bar{x}+t_nz_n))\| + \|h'(\bar{x})+h''(\bar{x})t_nz_n-h'(\bar{x}+t_nz_n)\|.$$

Together with

$$\|h'(\xi)-A_n\| \le \|h'(\xi)-h'(\bar{x}+t_nz_n)\| + \|h'(\bar{x}+t_nz_n)-A_n\|$$

and $\|\bar{x}+t_nz_n-\xi\| \le \max\{\|\bar{x}+t_nz_n-x'\|, \|\bar{x}+t_nz_n-x\|\}$ condition (3.4) follows with $\varphi'_n := t_n^{-1}(2\sup\{\|h'(\bar{x})+h''(\bar{x})(\eta-\bar{x})-h'(\eta)\| : \eta \in \bar{x}+t_n(z_n+R'\mathcal{B}_X)\}+\|h'(\bar{x}+t_nz_n)-A_n\|) = o(1)$ and $\gamma' = \|h''(\bar{x})\|$. Similarly, when $h'(\cdot)$ is Lipschitz continuous in $\bar{x}+\rho\mathcal{B}_X$, condition (3.4) holds with $\varphi'_n := t_n^{-1}\|h'(\bar{x}+t_nz_n)-A_n\| = o(1)$ and with γ' being the Lipschitz constant of $h'(\cdot)$. Thus the lemma is proved. □

Note that a sequence $(A_n) \subset L(X,Y)$ satisfying (3.2) can also exist if h is not continuously differentiable. For instance, consider a twice continuously differentiable function $\Psi : \mathbb{R} \to \mathbb{R}$ with $\Psi(t) = o(t^2)$ for $t \to 0$, let $\tilde{y} \in Y$ be arbitrarily chosen, and set $h(x) := \Psi(\|x-\bar{x}\|)\tilde{y}$. Then it is easy to show that condition (3.2) holds with $A_n = 0$ for all n, but in general $h(\cdot)$ will be only continuously differentiable provided $\|\cdot-\bar{x}\|^2$ is.

In addition, this example together with the second part of Theorem 2.2 shows that the conclusion of Theorem 3.1 automatically holds at a point $\bar{x}$ which is nondegenerate for the problem (P) and where condition (2.1) is satisfied. On the other hand, when $\bar{x}$ is degenerate for the problem (P) (i.e., $\operatorname{int}(h'(\bar{x})X - C) = \emptyset$), then, as a consequence of Theorem 2.4, the necessary condition (2.1) is automatically satisfied regardless of whether the point $\bar{x}$ is a local weak minimizer for the problem (P) or not. However, as we will see, condition (3.3) of Theorem 3.1 may fail for nonoptimal points $\bar{x}$. We will present corresponding examples in sections 4 and 5.

Further note that sequences $(z_n), (t_n), (\tau_n)$ satisfying the assumption (3.1) exist if and only if condition (2.2) does not hold, or, equivalently, $\bar{x}$ is not an essential local minimizer of second order. Thus the necessary condition (3.3) reduces the gap between the necessary and sufficient conditions of Theorem 3.1 in the degenerate case.

4. Second-order necessary conditions for certain directions. We will now analyze Theorem 3.1 for the special case of convergent sequences $(z_n) \to z$, and we will rewrite condition (3.3) in terms of first- and second-order derivatives of h and first- and second-order approximation sets for the convex set C. Since we deal with rather general sets C, there is an inherent nonsmoothness when building second-order approximation sets. Hence it seems to be quite natural to assume a similar amount of smoothness on h only.

DEFINITION 4.1. *Let E, F be normed spaces and let an element $z \in E$ be given.*

1. *Let $k : E \to F$ be a mapping and let $\bar{e} \in E$ so that k is differentiable at $\bar{e}$. We define the following second-order one-sided directional derivative to k at $\bar{e}$ with respect to z as*

$$k''(\bar{e};z) := \left\{ f \in F \,:\, \exists (t_n) \in \mathbb{T} \text{ such that } f = \lim_{n\to\infty} \frac{k(\bar{e}+t_nz)-k(\bar{e})-t_nk'(\bar{e})z}{t_n^2/2} \right\}.$$

Further, for given $\vec{t} = (t_n) \in \mathbb{T}$ we write

$$k''_{\vec{t}}(\bar{e};z) := \lim_{n\to\infty} \frac{k(\bar{e}+t_nz)-k(\bar{e})-t_nk'(\bar{e})z}{t_n^2/2},$$

when the limit on the right-hand side exists.

2. *Let S be a subset of F, let $A \in L(E,F)$ be a continuous linear operator, and let $\bar{f} \in S$. Then for $z \in E$ the* second-order compound tangent set *to S at $(\bar{f}, z)$ (with respect to A) and a sequence $\vec{t} = (t_n) \in \mathbb{T}$ is the set*

$$S''_{A,\vec{t}}(\bar{f};z) := \left\{ w \in Y \,:\, \exists (z_n) \to z \;\; \text{such that} \;\; d\left(\bar{f}+t_nAz_n+\frac{t_n^2}{2}w, S\right) = o(t_n^2) \right\}.$$

We also define

$$S''_A(\bar f, z) := \bigcup_{\vec t\in\mathbb{T}} S''_{A,\vec t}(\bar f; z),$$

which corresponds to the second-order compound tangent set introduced in [28] and which played a crucial role in [15].

In order to have $C''_{h'(\bar x),\vec t}(h(\bar x); z) \neq \emptyset$ it must necessarily hold that

$$d(h(\bar x) + t_n h'(\bar x)z, C) = d(h(\bar x) + t_n h'(\bar x)z_n, C) + o(t_n) = O(t_n^2) + o(t_n) = o(t_n)$$

for some sequence $(z_n) \to z$, implying $h'(\bar x)z \in T_C(h(\bar x))$, i.e., z belongs to the so-called *critical cone* $\mathcal{C}(\bar x)$ defined by

$$\mathcal{C}(\bar x) := \{z \in X \,:\, h'(\bar x)z \in T_C(h(\bar x))\}.$$

LEMMA 4.2. *For any element $z \in \mathcal{C}(\bar x)$ and any sequence $\vec t = (t_n) \in \mathbb{T}$ with $C''_{h'(\bar x),\vec t}(h(\bar x); z) \neq \emptyset$, the inclusions*

$$\mathrm{cl}\,(C''_{h'(\bar x),\vec t}(h(\bar x); z) + T_C(h(\bar x)) + \mathrm{Im}\, h'(\bar x)) \subset C''_{h'(\bar x),\vec t}(h(\bar x); z) \subset \mathrm{cl}\,(T_C(h(\bar x)) + \mathrm{Im}\, h'(\bar x))$$

hold. Moreover, $C''_{h'(\bar x),\vec t}(h(\bar x); z)$ is a closed convex set.

Proof. The fact that $C''_{h'(\bar x),\vec t}(h(\bar x); z)$ is a closed convex set follows easily from the definition. To show the inclusions let $y \in C''_{h'(\bar x),\vec t}(h(\bar x); z)$ be arbitrarily fixed. By the definition we can find sequences $(z_n) \to z$ and $(y_n) \to y$ such that $h(\bar x) + t_n h'(\bar x)z_n + \frac{1}{2}t_n^2 y_n \in C$ for each n and $y_n \in 2t_n^{-2}(C - h(\bar x)) + h'(\bar x)(-t_n^{-1}z_n) \subset T_C(h(\bar x)) + \mathrm{Im}\, h'(\bar x)$ follows. Hence $y \in \mathrm{cl}\,(T_C(h(\bar x)) + \mathrm{Im}\, h'(\bar x))$. Now consider arbitrary elements $w \in T_C(h(\bar x))$ and $v = h'(\bar x)s \in \mathrm{Im}\, h'(\bar x)$. Then we can find a convergent sequence $(w_n) \to w$ such that $h(\bar x) + t_n w_n \in C$ for all n. For all n sufficiently large we have $t_n < 2$ and by using the convexity of C we conclude

$$\begin{aligned}\frac{t_n}{2}(h(\bar x) + t_n w_n) + \left(1 - \frac{t_n}{2}\right)\left(h(\bar x) + t_n h'(\bar x)z_n + \frac{1}{2}t_n^2 y_n\right)\\ = h(\bar x) + t_n h'(\bar x)\left(z_n - \frac{t_n}{2}(z_n + s)\right) + \frac{t_n^2}{2}(y_n + w_n + v) \in C.\end{aligned}$$

Since the sequence $(z_n - \frac{t_n}{2}(z_n + s))$ converges to z we obtain $y + w + v = \lim_n y_n + w_n + v \in C''_{h'(\bar x),\vec t}(h(\bar x); z)$, and since $C''_{h'(\bar x),\vec t}(h(\bar x); z)$ is closed, the proposed inclusion follows. □

LEMMA 4.3. *Suppose Assumption 1 is satisfied. Let $(z_n) \to z$ be a convergent sequence in X and let $\vec t = (t_n) \in \mathbb{T}$, such that $h''_{\vec t}(\bar x; z)$ exists. Then*

$$h''_{\vec t}(\bar x; z) = \lim_{n\to\infty} \frac{h(\bar x + t_n z_n) - h(\bar x) - t_n h'(\bar x)z_n}{t_n^2/2}$$

also holds. Moreover, there exists a sequence $\tau_n \to 0$ such that (3.1) holds if and only if $h''_{\vec t}(\bar x; z) \in C''_{h'(\bar x),\vec t}(h(\bar x); z)$.

Proof. Using Assumption 1, the first assertion follows immediately from the estimate

$$\|h(\bar x + t_n z_n) - h(\bar x + t_n z) - t_n h'(\bar x)(z_n - z)\| \le \eta t_n^2 \max\{\|z_n\|, \|z\|\}\|z_n - z\| = o(t_n^2).$$

Now assume $h''_{\bar t}(\bar x;z)\in C''_{h'(\bar x),\bar t}(h(\bar x);z)$. By the definition, there exists a sequence $(z'_n)\to z$ such that $d(h(\bar x)+t_nh'(\bar x)z'_n+(t_n^2/2)h''_{\bar t}(\bar x;z),C)=o(t_n^2)$. Hence,

$$\begin{aligned}&d(h(\bar x+t_nz_n)+t_nh'(\bar x)(z'_n-z_n),C)\\&\quad=d\left(h(\bar x)+t_nh'(\bar x)z_n+\frac{t_n^2}{2}h''_{\bar t}(\bar x;z)+t_nh'(\bar x)(z'_n-z_n),C\right)+o(t_n^2)=o(t_n^2)\end{aligned}$$

and (3.1) follows from Theorem 2.4 with $\tau_n=\|z'_n-z_n\|$. Now let $(\tau_n)\in\mathbb{T}$ be a sequence such that (3.1) holds. Then, as in the proof of Theorem 3.1, we can find some sequence $(z'_n)\to z$ with $z'_n=z_n+\tau_nt_n\delta z_n$, $\delta z_n\in\mathcal{B}_X$, such that

$$d(h(\bar x+t_nz'_n),C)=d\left(h(\bar x)+t_nh'(\bar x)z'_n+\frac{t_n^2}{2}h''_{\bar t}(\bar x;z),C\right)+o(t_n)^2=o(t_n^2),$$

showing $h''_{\bar t}(\bar x;z)\in C''_{h'(\bar x),\bar t}(h(\bar x);z)$. ∎

The second-order derivative $h''_{\bar t}(\bar x;z)$ is useful for building second-order approximations of h in the direction z. But we also need another type of second-order derivatives, namely, in the sense of first-order approximations of first derivatives. We know from the discussion following Theorem 3.1 that if h is sufficiently smooth near $\bar x$, then for given sequences $(z_n)\subset\mathcal{S}_X$ and $(t_n)\in\mathbb{T}$ a sequence of linear operators $(A_n)\subset L(X,Y)$ satisfies condition (3.2) if and only if $A_n=h'(\bar x+t_nz_n)+o(t_n)$ holds. We use condition (3.2) to define this other type of second-order derivative without assuming existence of $h'(\cdot)$ near $\bar x$.

LEMMA 4.4. *Let E,F be normed spaces, let $k:E\to F$ be a mapping, and let $\bar e\in E$ such that k is differentiable at $\bar e$. Further let an element $z\in E$ and a sequence $(t_n)\in\mathbb{T}$ be given. Then there exists at most one continuous linear operator $K\in L(E,F)$ such that, together with some positive scalars $\tilde\gamma$, $\tilde R$ and some sequence $(\tilde\varphi_n)\in\mathbb{T}$, one has*

$$\begin{aligned}&\|k(e')-k(e)-(k'(\bar e)+t_nK)(e'-e)\|\\&\qquad\le\big(\tilde\varphi_nt_n+\tilde\gamma\max\{\|\bar e+t_nz-e'\|,\|\bar e+t_nz-e\|\}\big)\|e'-e\|\end{aligned}\tag{4.1}$$

for all $e,e'\in\bar e+t_n(z+\tilde R\mathcal{B}_E)$ and all n.

Proof. We prove the lemma by contraposition. Assume that there exist two continuous linear operators $K_1\ne K_2$ satisfying (4.1) with parameters $\tilde\gamma_1,\tilde R_1,(\tilde\varphi_{1n})$ and $\tilde\gamma_2,\tilde R_2,(\tilde\varphi_{2n})$, respectively. Set $\tilde\gamma:=\max\{\tilde\gamma_1,\tilde\gamma_2\}$, $\tilde\varphi_n:=\max\{\tilde\varphi_{1n},\tilde\varphi_{2n}\}$ for all n, and $\tilde R=\min\{\tilde R_1,\tilde R_2\}$. By the triangle inequality we obtain

$$\begin{aligned}\|t_n(K_1-K_2)(e'-e)\|&\le\|k(e')-k(e)-(k'(\bar e)+t_nK_1)(e'-e)\|\\&\quad+\|k(e')-k(e)-(k'(\bar e)+t_nK_2)(e'-e)\|\\&\le2\big(\tilde\varphi_nt_n+\tilde\gamma\max\{\|\bar e+t_nz-e'\|,\|\bar e+t_nz-e\|\}\big)\|e'-e\|\end{aligned}$$

for all $e,e'\in\bar e+t_n(z+\tilde R\mathcal{B}_E)$ and all n. Let $d\in E$ denote a direction with $(K_1-K_2)d\ne0$ and $\|d\|\le R'$. Applying the above estimate successively with $e=\bar e+t_nz$, $e'=e+\frac{t_n}{n}d$ for each n we obtain

$$\left\|t_n(K_1-K_2)\frac{t_n}{n}d\right\|\le2\left(\tilde\varphi_nt_n+\tilde\gamma\left\|\frac{t_n}{n}d\right\|\right)\left\|\frac{t_n}{n}d\right\|=2\frac{t_n^2}{n}\left(\tilde\varphi_n+\frac{\tilde\gamma\|d\|}{n}\right)\|d\|.$$

Dividing by $\frac{t_n^2}{n}$ and passing to the limit yields $\|(K_1-K_2)d\|\le0$, a contradiction. ∎

DEFINITION 4.5. *Under the assumptions of Lemma* 4.4*, if the unique continuous linear operator* $K \in L(E,F)$ *satisfying condition* (4.1) *exists, we will denote it by* $(k')'_{\vec{t}}(\bar{e};z)$.

If k is differentiable near $\bar{e}$, then $\|k'(\bar{e}+t_nz)-(k'(\bar{e})+t_n(k')'_{\vec{t}}(\bar{e};z))\| = o(t_n)$ follows easily from condition (4.1). Thus, $(k')'_{\vec{t}}(\bar{e};z)$ is an element of the so-called *contingent derivative*, also called *graphical derivative* or *Bouligand derivative* (see, e.g., [24]), which in our case is given by

$$Ck'(\bar{e})z := \left\{ K \in L(E,F) \,:\, \exists (t_n) \in \mathbb{T}, z_n \to z \text{ with } \; K = \lim_{n\to\infty} \frac{k'(\bar{e}+t_nz_n)-k'(\bar{e})}{t_n} \right\}.$$

If $k'(\cdot)$ is Lipschitz continuous near $\bar{e}$, using arguments similar to those in Lemma 3.2 we can conclude that

$$\bigcup_{\vec{t}\in T} (k')'_{\vec{t}}(\bar{e};z) = Ck'(\bar{e})z,$$

and if k is twice Fréchet differentiable at $\bar{e}$, then $(k')'_{\vec{t}}(\bar{e};z) = k''(\bar{e})z$ for all $\vec{t} \in \mathbb{T}$ holds.

In general, when k is not twice differentiable at $\bar{e}$, we can have $(k')'_{\vec{t}}(\bar{e};z)z \neq k''_{\vec{t}}(\bar{e};z)$. However, $(k')'_{\vec{t}}(\bar{e};z)$ acts like a derivative for the mapping $\frac{1}{2}k''_{\vec{t}}(\bar{e};\cdot)$ at z for given $\vec{t}\in\mathbb{T}$. Indeed, from condition (4.1) the estimate

$$\left\| \frac{1}{2}k''_{\vec{t}}(\bar{e};s) - \frac{1}{2}k''_{\vec{t}}(\bar{e};z) - (k')'_{\vec{t}}(\bar{e};z)(s-z) \right\| \le \tilde{\gamma}\|s-z\|^2,$$

being valid for all $s \in z + \tilde{R}\mathcal{B}_X$ such that $k''_{\vec{t}}(\bar{e};s)$ exists, easily follows.

When $(h')'_{\vec{t}}(\bar{x};z)$ exists for some $z \in X$, $\vec{t}\in\mathbb{T}$, then it is easy to see that for any convergent sequence $(z_n) \to z$ condition (3.2) holds with $A_n = h'(\bar{x}) + t_n(h')'_{\vec{t}}(\bar{x};z)$, $\varphi'_n = \tilde{\varphi}_n + \tilde{\gamma}\|z_n - z\|$, $\gamma' = \tilde{\gamma}$, $R' = \tilde{R}/2$ for all n sufficiently large, such that $\|z_n - z\| \le \tilde{R}/2$.

We are now in a position to state the main result of this section. We state this result under the following technical assumption, which will be analyzed separately.

ASSUMPTION 2. *Each sequence* $(y^*_n) \subset \mathcal{S}_{Y^*}$ *with*

$$\lim_{n\to\infty} t_n^{-1}(\langle y^*_n, h(\bar{x})\rangle - \sigma_C(y^*_n)) = \lim_{n\to\infty} t_n^{-1}\|(h'(\bar{x}) + t_n(h')'_{\vec{t}}(\bar{x};z))^*y^*_n\| = 0 \tag{4.2}$$

has at least one weak- accumulation point which is not equal to* 0.

THEOREM 4.6. *Suppose that* $\bar{x}$ *is a local weak minimizer for problem* (P)*, that Assumption* 1 *holds, and that we are given an element* $z \in \mathcal{C}(\bar{x}) \cap \mathcal{S}_X$ *and a sequence* $(t_n) = \vec{t} \in \mathbb{T}$ *such that the second-order directional derivatives* $h''_{\vec{t}}(\bar{x};z)$ *and* $(h')'_{\vec{t}}(\bar{x};z)$ *exist, the inclusion* $h''_{\vec{t}}(\bar{x};z) \in C''_{h'(\bar{x}),\vec{t}}(h(\bar{x});z)$ *holds, and Assumption* 2 *is satisfied. Then there exist a multiplier* $\bar{y}^* \in \Lambda_{FJ}$ *and an element* $\bar{\mu}^* \in (\operatorname{Ker} h'(\bar{x}))^\perp$ *such that*

$$(h')'_{\vec{t}}(\bar{x};z)^*\bar{y}^* + \bar{\mu}^* = 0, \tag{4.3}$$

and for each pair (s,y) *with* $s \in \mathcal{C}(\bar{x})$ *and* $y \in C''_{h'(\bar{x}),\vec{t}}(h(\bar{x});s)$ *one has*

$$\langle \bar{\mu}^*, z-s\rangle + \frac{1}{2}\langle \bar{y}^*, h''_{\vec{t}}(\bar{x};z) - y\rangle \ge 0. \tag{4.4}$$

Proof. From the preceding discussion we know that the assumptions of Theorem 3.1 hold with $z_n := z$, $A_n := h'(\bar{x}) + t_n(h')'_{\bar{t}}(\bar{x}; z)$, and some sequence $(\tau_n) \in \mathbb{T}$ given by Lemma 4.3. Hence, applying Theorem 3.1 with $T = 1$ we have

$$\liminf_{n\to\infty} \sup_{y^*\in\mathcal{S}_{Y^*}} \left\{ \frac{\langle y^*, h(\bar{x}+t_nz)\rangle - \sigma_C(y^*)}{t_n^2} - \frac{1}{t_n}\|(h'(\bar{x}) + t_n(h')'_{\bar{t}}(\bar{x}; z))^* y^*\| \right\} \geq 0.$$

Now, for each n we can find some $y_n^* \in \mathcal{S}_{Y^*}$ approaching the supremum sufficiently accurate, such that

$$\liminf_{n\to\infty} \left\{ \frac{\langle y_n^*, h(\bar{x}+t_nz)\rangle - \sigma_C(y_n^*)}{t_n^2} - \frac{1}{t_n}\|(h'(\bar{x}) + t_n(h')'_{\bar{t}}(\bar{x}; z))^* y_n^*\| \right\} \geq 0.$$

Since $\limsup_{n\to\infty} t_n^{-2}\hat{d}_C(h(\bar{x}+t_nz), h'(\bar{x}), \tau_nt_n) \leq 0$ we also have

$$\begin{aligned}
0 &\geq \limsup_{n\to\infty} \sup_{y^*\in\mathcal{S}_{Y^*}} \left\{ \frac{\langle y^*, h(\bar{x}+t_nz)\rangle - \sigma_C(y^*)}{t_n^2} - \frac{\tau_n}{t_n}\|(h'(\bar{x})^* y^*\| \right\} \\
&= \limsup_{n\to\infty} \sup_{y^*\in\mathcal{S}_{Y^*}} \left\{ \frac{\langle y^*, h(\bar{x}+t_nz)\rangle - \sigma_C(y^*)}{t_n^2} - \frac{\tau_n}{t_n}\|h'(\bar{x})^* y^*\| + \tau_n\|(h')'_{\bar{t}}(\bar{x}; z)^* y^*\| \right\} \\
&\geq \limsup_{n\to\infty} \sup_{y^*\in\mathcal{S}_{Y^*}} \left\{ \frac{\langle y^*, h(\bar{x}+t_nz)\rangle - \sigma_C(y^*)}{t_n^2} - \frac{\tau_n}{t_n}\|(h'(\bar{x}) + t_n(h')'_{\bar{t}}(\bar{x}; z))^* y^*\| \right\}
\end{aligned}$$

and $\lim_{n\to\infty} t_n^{-1}\|(h'(\bar{x}) + t_n(h')'_{\bar{t}}(\bar{x}; z))^* y_n^*\| = 0$ follows. In particular we obtain

$$\mu_n^* := \frac{h'(\bar{x})^* y_n^*}{t_n} = O(1), \tag{4.5}$$

$$\lim_{n\to\infty} t_n^{-1}(h'(\bar{x}) + t_n(h')'_{\bar{t}}(\bar{x}; z))^* y_n^* = \lim_{n\to\infty} (h')'_{\bar{t}}(\bar{x}; z)^* y_n^* + \mu_n^* = 0, \tag{4.6}$$

$$\liminf_{n\to\infty} \frac{\langle y_n^*, h(\bar{x}+t_nz)\rangle - \sigma_C(y_n^*)}{t_n^2} \geq 0. \tag{4.7}$$

Because of (4.7) and the relation $h(\bar{x}+t_nz) = h(\bar{x}) + t_nh'(\bar{x})z + \frac{t_n^2}{2}h''_{\bar{t}}(\bar{x}; z) + o(t_n^2)$ we have

$$\liminf_{n\to\infty} \left\{ \frac{\langle y_n^*, h(\bar{x})\rangle - \sigma_C(y_n^*)}{t_n^2} + \frac{\langle y_n^*, h'(\bar{x})z\rangle}{t_n} + \frac{1}{2}\langle y_n^*, h''_{\bar{t}}(\bar{x}; z)\rangle \right\} \geq 0.$$

Together with $\langle y_n^*, h'(\bar{x})z\rangle = t_n\langle \mu_n^*, z\rangle = O(t_n)$ and $\langle y_n^*, h(\bar{x})\rangle - \sigma_C(y_n^*) \leq 0$ we obtain

$$\sigma_C(y_n^*) - \langle y_n^*, h(\bar{x})\rangle = O(t_n^2). \tag{4.8}$$

Together with (4.6) we may conclude from Assumption 2 that the sequence (y_n^*) has a nonzero accumulation point $\tilde{y}^*$. Then there exists some element $\tilde{y} \in Y$ with $\langle \tilde{y}^*, \tilde{y}\rangle = 1$, and we can choose a subsequence $(y_{k_n}^*)$ such that $\langle y_{k_n}^*, \tilde{y}\rangle \to 1$. $(y_{k_n}^*, \mu_{k_n}^*)$ is a bounded sequence in $Y^* \times X^* = (Y \times X)^*$ and by the Alaoglu–Bourbaki theorem at least one weak-$*$ accumulation point, say $(\bar{y}^*, \bar{\mu}^*)$, exists. Of course, $\bar{y}^*$ is also a weak-$*$ accumulation point of the sequence $(y_{k_n}^*)$. Hence $\langle \bar{y}^*, \tilde{y}\rangle = 1$, implying $\bar{y}^* \neq 0$. Note that $(\bar{y}^*, \bar{\mu}^*)$ is also a weak-$*$ accumulation point of the entire sequence (y_n^*, μ_n^*). Since $\mu_n^* \in \operatorname{Im} h'(\bar{x})^* \subset (\operatorname{Ker} h'(\bar{x}))^\perp$ and the annihilator $(\operatorname{Ker} h'(\bar{x}))^\perp$ is weakly-$*$ closed in X^*, we have $\bar{\mu}^* \in (\operatorname{Ker} h'(\bar{x}))^\perp$. Further, (4.3) follows easily from (4.6).

Now let us show $\bar{y}^* \in \Lambda_{FJ}$. Since $\sigma_C(\cdot) - \langle \cdot, h(\bar{x})\rangle$ is weakly-$*$ lower semicontinuous, we obtain $\sigma_C(\bar{y}^*) - \langle \bar{y}^*, h(\bar{x})\rangle \leq 0$ from condition (4.8). Because of $h(\bar{x}) \in C$

we also have $\sigma_C(\bar{y}^*) - \langle \bar{y}^*, h(\bar{x})\rangle \geq 0$, showing $\bar{y}^* \in N_C(h(\bar{x}))$. From condition (4.5) it follows that $h'(\bar{x})y_n^* \to 0$ and, consequently, $h'(\bar{x})^*\bar{y}^* = 0$. Hence $\bar{y}^* \in \Lambda_{FJ}$. It remains to show that (4.4) holds. Let the pair $(s, y) \in \mathcal{C}(\bar{x}) \times C''_{h'(\bar{x}),\vec{t}}(h(\bar{x}); s)$ be arbitrarily fixed. Then by the definitions of the support function σ_C and the set $C''_{h'(\bar{x}),\vec{t}}(h(\bar{x}); s)$ we have $\sigma_C(y_n^*) \geq \langle y_n^*, h(\bar{x}) + t_n h'(\bar{x})s_n + \frac{t_n^2}{2}y\rangle + o(t_n^2)$ for some sequence $(s_n) \to s$. Together with condition (4.7) and the second-order expansion for $h(\bar{x} + t_n z)$ it follows that

$$0 \leq \liminf_{n\to\infty} \frac{\langle y_n^*, h(\bar{x}) + t_n h'(\bar{x})z + \frac{t_n^2}{2}h''_{\vec{t}}(\bar{x}; z)\rangle - \langle y_n^*, h(\bar{x}) + t_n h'(\bar{x})s_n + \frac{t_n^2}{2}y\rangle}{t_n^2}$$

$$= \liminf_{n\to\infty} \left\{ \langle \mu_n^*, z - s_n\rangle + \frac{1}{2}\langle y_n^*, h''_{\vec{t}}(\bar{x}; z) - y\rangle \right\} \leq \langle \bar{\mu}^*, z - s\rangle + \frac{1}{2}\langle \bar{y}^*, h''_{\vec{t}}(\bar{x}; z) - y\rangle,$$

and this completes the proof. □

Of course, Assumption 2 holds when Y is finite-dimensional. But there are also other situations when this assumption holds.

DEFINITION 4.7. *Let E, F be Banach spaces, let $S \subset F$ be a closed convex subset of F, let $k : E \to F$ be a mapping, which is differentiable at $\bar{e} \in E$, with $k(\bar{e}) \in S$, and let the multifunction $\Psi : E \rightrightarrows F$ be given by $\Psi(e) := k(e) - S$. Then k is said to be 2-nondegenerate at the point $\bar{e}$ in the direction $z \in E$ with respect to the set C and the sequence $\vec{t} \in \mathbb{T}$ if the following conditions are satisfied:*

1. *$(k')'_{\vec{t}}(\bar{e}; z)$ exists.*
2. *The set $k(\bar{e}) + k'(\bar{e})E - S$ has nonempty relative interior.*
3. *The interior of the set $KE - (S - k(\bar{E})) \times \{0\}$ is nonempty in $Q \times F/Q$, where $Q := \operatorname{aff}(k(\bar{e}) + k'(\bar{e})E - S) \subset F$ is the affine hull of the set $k(\bar{e}) + k'(\bar{e})E - S$, the continuous linear operator $K : E \to Q \times F/Q$ is given by $Ks := (k'(\bar{e})s, \pi(k')'_{\vec{t}}(\bar{e}; z)s)$, and π denotes the quotient map from F onto the quotient space F/Q.*

In what follows let Q denote the affine hull of the set $h(\bar{x}) + h'(\bar{x})X - C$ and let π denote the quotient map from Y onto the quotient space Y/Q.

THEOREM 4.8. *Let the mapping h be 2-nondegenerate at $\bar{x}$ in the direction $z \in X$ with respect to C and the sequence $\vec{t} \in \mathbb{T}$. Then it is sufficient for Assumption 2 to hold that either the quotient space Y/Q or the subspace $\operatorname{Im} h'(\bar{x}) \cap \operatorname{aff}(C - h(\bar{x}))$ is finite-dimensional.*

Proof. Let $(y_n^*) \subset \mathcal{S}_{Y^*}$ be a sequence satisfying condition (4.2). In order to prove the theorem we have to show that at least one nonzero weak-$*$ accumulation point of the sequence (y_n^*) exists. Since Q is a closed subspace of the Banach space Y, the quotient space $P := Y/Q$ and hence also $Q \times P$ are Banach spaces. Let $H : X \to Q \times P$ be the continuous linear operator according to Definition 4.7, i.e., $Hs = (h'(\bar{x})s, \pi(h')'_{\vec{t}}(\bar{x}; z)s)$ for all s. Now choose $(q, p) \in Q \times P$, $x \in X$, and $\bar{q} \in \hat{C} := C - h(\bar{x}) \subset Q$ such that $(q, p) = Hx - (\bar{q}, 0) \in \operatorname{int}(HX - \hat{C} \times \{0\})$. Application of the generalized open mapping theorem (see [31, Theorem 1]) yields $(q, p) \in \operatorname{int}(H(x + \mathcal{B}_X) - \hat{C} \times \{0\})$ and it follows that

$$\rho\mathcal{B}_{Q\times P} \subset \operatorname{int}((\bar{q}, 0) + H\mathcal{B}_X - \hat{C} \times \{0\}) \tag{4.9}$$

for some $\rho > 0$. Using Theorem 2.4 we obtain that

$$\sup_{(q^*,p^*)\in\mathcal{S}_{(Q\times P)^*}} \left\{ \langle q^*, \bar{q}\rangle - \sigma_{\hat{C}}(q^*) - \|H^*(q^*, p^*)\| \right\} \leq -\rho$$

and therefore

$$\langle q^*, \bar{q}\rangle - \sigma_{\hat{C}}(q^*) - \|H^*(q^*,p^*)\| \le -\rho(\|q^*\| + \|p^*\|)\ \forall (q^*,p^*) \in Q^* \times P^*. \tag{4.10}$$

Now let the linear operators $A \in L(X,Q)$ and $B \in L(X,P)$ be given by $As = h'(\bar{x})s$ and $Bs = \pi(h')'_{\bar{t}}(\bar{x};z)s$, respectively. Further let $i_Q : Q \to Y$ denote the natural embedding from Q into Y. Consequently, $h'(\bar{x}) = i_Q \circ A$. For each n, let $q_n^* := i_Q^* y_n^* \in Q^*$ be the restriction of the linear functional y_n^* to Q. By the Hahn–Banach theorem we can extend the linear form $q_n^* \in Q^*$ to a linear functional $y_{Q,n}^*$ over Y such that $\|y_{Q,n}^*\| = \|q_n^*\|$. Setting $y_{P,n}^* := y_n^* - y_{Q,n}^*$ we have $i_Q^* y_{P,n}^* = i_Q^* y_n^* - i_Q^* y_{Q,n}^* = 0$, i.e., $y_{P,n}^*$ belongs to the annihilator $Q^\perp := \{y^* \in Y^* : \langle y^*, y\rangle = 0 \text{ for all } y \in Q\}$ of the subspace $Q \subset Y$. The mapping $\pi^* : P^* \to Y^*$ is isometric and allows us to identify the dual space $P^* = (Y/Q)^*$ with $Q^\perp$. For each n let $p_n^* \in P^*$ denote the linear functional uniquely given by the relation $\pi^* p_n^* = y_{P,n}^*$. Finally, let $\bar{y}_Q := i_Q \bar{q}$. Then we have

$$\langle y_n^*, \bar{y}_Q\rangle = \langle q_n^*, \bar{q}\rangle, \tag{4.11}$$

$$\sigma_{\hat{C}}(q_n^*) = \sigma_{i_Q(\hat{C})}(y_n^*) = \sigma_C(y_n^*) - \langle y_n^*, h(\bar{x})\rangle, \tag{4.12}$$

$$\begin{aligned}(h'(\bar{x})^* + t_n(h')'_{\bar{t}}(\bar{x};z))^* y_n^* &= A^* q_n^* + t_n B^* p_n^* + t_n(h')'_{\bar{t}}(\bar{x};z)^* y_{Q,n}^* \\ &= H^*(q_n^*, t_n p_n^*) + t_n(h')'_{\bar{t}}(\bar{x};z)^* y_{Q,n}^*,\end{aligned} \tag{4.13}$$

and (4.10) implies

$$\begin{aligned}&\langle y_n^*, \bar{y}_Q\rangle + \langle y_n^*, h(\bar{x})\rangle - \sigma_C(y_n^*) - \|(h'(\bar{x})^* + t_n(h')'_{\bar{t}}(\bar{x};z))^* y_n^*\| \\ &\qquad \le -\rho(\|q_n^*\| + t_n\|p_n^*\|) + t_n\|(h')'_{\bar{t}}(\bar{x};\bar{z})^* y_{Q,n}^*\|.\end{aligned} \tag{4.14}$$

Now let us assume that $\limsup_{n\to\infty} \|q_n^*\| > 0$. By passing to a subsequence if necessary we may also assume that $\liminf_{n\to\infty} \|q_n^*\| = \epsilon > 0$. The sequence (y_n^*) has at least one weak-$*$ accumulation point $\bar{y}^*$ by the Alaoglu–Bourbaki theorem. Since $t_n \to 0$ and the sequence (y_n^*) satisfies condition (4.2) we obtain from condition (4.14) that $\langle \bar{y}^*, \bar{y}_Q\rangle \le -\rho\epsilon < 0$. Hence, $\bar{y}^* \ne 0$ and the theorem is proved in the case $\limsup_{n\to\infty} \|q_n^*\| > 0$.

Now let us assume that $\limsup_{n\to\infty} \|q_n^*\| = 0$. If we denote by $\bar{p}^*$ an arbitrary weak-$*$ accumulation point of the sequence (p_n^*), then $\bar{y}^* := \pi^* \bar{p}^*$ is a weak-$*$ accumulation point both of the sequence $(y_{P,n}^*)$ and the sequence (y_n^*), the latter because of $\|y_{Q,n}^*\| = \|y_n^* - y_{P,n}^*\| \to 0$. Further, $\bar{y}^* \ne 0$ if and only if $\bar{p}^* \ne 0$. Since $\|y_{P,n}^*\| = \|p_n^*\| \to 1$, the assertion of the theorem now follows immediately in the case that P and hence also P^* are finite-dimensional spaces.

It remains to prove the theorem in the case in which the subspace $W := \operatorname{Im} h'(\bar{x}) \cap \operatorname{aff}(C - h(\bar{x}))$ is finite-dimensional. To do this we will first show the inclusion

$$\rho\mathcal{B}_{Q\times P} \subset (\bar{q},0) + \left(\frac{A}{t}, B\right)\mathcal{B}_X - \left(\hat{C} + \frac{\rho + \|A\|}{t}\mathcal{B}_W\right) \times \{0\}\ \forall t \in \left]0, \frac{\rho}{\|A\|}\right[. \tag{4.15}$$

Let the element $(q,p) \in \rho\mathcal{B}_{Q\times P}$ and the scalar $t \in]0,\ \|A\|^{-1}\rho[$ be arbitrarily fixed. We observe that condition (4.9) together with $\bar{q} \in \hat{C}$ implies $\gamma\rho\mathcal{B}_{Q\times P} \subset (\bar{q},0) + \gamma H\mathcal{B}_X - \hat{C} \times \{0\}$ for all $\gamma \in]0,1]$. Hence we can find some $x_1 \in \rho^{-1}\|(p,q)\|\mathcal{B}_X$ and some $c_1 \in \hat{C}$ such that $(q,p) = (\bar{q} + Ax_1 - c_1, Bx_1)$. We have $\|t(q - \bar{q} + c_1)\| = \|tAx_1\| \le t\|A\|\|x_1\| \le \|(q,p)\|$, and $\|(t(q-\bar{q}+c_1),p)\| = \max\{\|t(q-\bar{q}+c_1)\|, \|p\|\} \le \|(q,p)\| \le \rho$

follows. Consequently, we can find some $x_2 \in \rho^{-1}\|(p,q)\|\mathcal{B}_X$ and some $c_2 \in \hat{C}$ with $(t(q-\bar{q}+c_1),p) = (\bar{q}+Ax_2-c_2, Bx_2)$. Thus

$$(q,p) = \left(\bar{q} + \frac{\bar{q}-c_2}{t} + \frac{1}{t}Ax_2 - c_1, Bx_2\right)$$

and $\bar{q}-c_2 = A(tx_1-x_2) \in W$. Moreover, we have

$$\|\bar{q}-c_2\| \le t\|Ax_1\| + \|A\|\|x_2\| \le \|(p,q)\|\left(1+\frac{\|A\|}{\rho}\right) \le \rho + \|A\|.$$

Therefore, $(p,q) \in (\bar{q},0) - (\hat{C} + t^{-1}(\rho+\|A\|)\mathcal{B}_W) \times \{0\} + (t^{-1}A, B)\mathcal{B}_X$ and, since $(p,q) \in \rho\mathcal{B}_{Q\times P}$ has been chosen arbitrarily, the inclusion (4.15) follows. Now, by Theorem 2.4 we obtain

$$\sup_{(q^*,p^*)\in\mathcal{S}^*_{Q\times P}} \left\{\langle q^*,\bar{q}\rangle - \sigma_{\hat{C}}(q^*) - (\rho+\|A\|)\frac{\sigma_{\mathcal{B}_W}(q^*)}{t} - \left\|A^*\frac{q^*}{t} + B^*p^*\right\|\right\} \le -\rho.$$

Consequently, for all n sufficiently large such that $t_n\|A\| < \rho$ we have

$$\langle q_n^*,\bar{q}\rangle - \sigma_{\hat{C}}(q_n^*) - (\rho+\|A\|)\frac{\sigma_{\mathcal{B}_W}(q_n^*)}{t_n} - \left\|A^*\frac{q_n^*}{t_n} + B^*p_n^*\right\| \le -\rho(\|q_n^*\| + \|p_n\|^*).$$

Taking into account conditions (4.2), (4.11)–(4.13), and $\|q_n^*\| = \|y^*_{Q,n}\| \to 0$ we obtain

$$\lim_{n\to\infty}\langle q_n^*,\bar{q}\rangle = 0,\ \lim_{n\to\infty}\sigma_{\hat{C}}(q_n^*) = 0,\ \lim_{n\to\infty}\left\|A^*\frac{q_n^*}{t_n} + B^*p_n^*\right\| = 0.$$

Since we also have $\|p_n^*\| \to 1$,

$$\liminf_{n\to\infty}\frac{\sigma_{\mathcal{B}_W}(q_n^*)}{t_n} \ge \frac{\rho}{\rho+\|A\|}$$

follows. Now, if W is finite-dimensional, and extracting if necessary a subsequence, there exists some $\tilde{x} \in X$ such that $\bar{w} := A\tilde{x} \in \mathcal{B}_W$ satisfies

$$\langle q_n^*,\bar{w}\rangle/t_n \ge \frac{\rho}{2(\rho+\|A\|)} := \tilde{\rho} > 0$$

for all n. Let $\bar{p}$ denote an arbitrary weak-$*$ accumulation point of the sequence (p_n^*). Extracting if necessary a subsequence, we have

$$\begin{aligned}\langle \bar{p}^*, B\tilde{x}\rangle &= \lim_{n\to\infty}\langle p_n^*, B\tilde{x}\rangle = \lim_{n\to\infty}\langle B^*p_n^*,\tilde{x}\rangle = -\lim_{n\to\infty}\left\langle A^*\frac{q_n^*}{t_n},\tilde{x}\right\rangle \\ &= -\lim_{n\to\infty}\left\langle\frac{q_n^*}{t_n},\bar{w}\right\rangle \le -\tilde{\rho},\end{aligned}$$

and $\bar{p}^* \ne 0$ follows. Then $\bar{y}^* = \pi^*\bar{p}^*$ is a nonzero weak-$*$ accumulation point of the sequence (y_n^*), and this completes the proof. □

Let us discuss the assumptions of Theorem 4.8 in further detail. In our case, the set C has the form $(-L)\times K$ with $\operatorname{int} L \ne \emptyset$, and so the assumption that $h(\cdot)$ is 2-nondegenerate in direction z with respect to C and $\vec{t}$ can be reduced to an assumption on the constraints, namely, that $g(\cdot)$ is 2-nondegenerate in direction z with respect to

K and $\vec{t}$. To be a little bit more general, let us consider the case when the Banach space Y and the set C can be decomposed in the form $C = C_1 \times C_2 \subset Y_1 \times Y_2 = Y$ with $\operatorname{int} C_1 \neq \emptyset$. Similarly, we denote by h_1, h_2, respectively, $(h_1')'_{\vec{t}}(\bar{x}; z), (h_2')'_{\vec{t}}(\bar{x}; z)$, the components of h, respectively, $(h')'_{\vec{t}}(\bar{x}; z)$. Then there holds $Q = Y_1 \times Q_2$ with $Q_2 = \operatorname{aff}(h_2(\bar{x}) + h_2'(\bar{x})X - C_2)$, and it follows that $h(\cdot)$ is 2-nondegenerate in direction z and with respect to C and $\vec{t}$ if and only if $h_2(\cdot)$ is of such a kind with respect to C_2.

Finally let us mention that the remaining assumption of Theorem 4.8, i.e., that either the quotient space Y/Q or the space $\operatorname{Im} h'(\bar{x}) \cap \operatorname{aff}(C - h(\bar{x}))$ is finite-dimensional, is satisfied in a variety of cases, e.g., when X or Y is finite-dimensional, or in the case of the scalar mathematical programming problem. Further, we know a lot of special cases where this assumption can be replaced, but we do not want to go down to the last detail here.

Example 4.1. Consider the problem

$$\min_{x \in \mathbb{R}^2} f(x) := x_1 - x_2^2$$
$$\text{subject to } g(x) := \begin{pmatrix} x_1 - \frac{1}{4}x_2^2 \\ x_1 x_2 \\ x_1 + \frac{1}{2}x_2^2 + \psi(x_2) \end{pmatrix} \in K,$$

where $K := \{(v_1, v_2, v_1 + v_2) : v_1 \geq |v_2|^{\frac{3}{2}}\}$ and

$$\psi(\xi) := \begin{cases} \xi^2 \sin(\ln|\xi|), & \xi \neq 0, \\ 0, & \xi = 0. \end{cases}$$

Note that ψ is not twice differentiable at 0; however, Assumption 1 is fulfilled. We show now that $\bar{x} := (0, 0)$ is not a local minimizer of this problem.

Define the set $C := \{(c_0, c_1, c_2, c_3) \colon c_0 \leq 0, (c_1, c_2, c_3) \in K\}$ and the mapping $h := (f - f(\bar{x}), g)$ according to (1.5). The set Λ_{FJ} consists of all multipliers $y_0^*, y_1^*, y_2^*, y_3^*, y_4^*$, not all 0, such that

$$y_0^* + y_1^* + y_3^* = 0, \quad y_0^* \geq 0, \quad y_1^* + y_3^* \leq 0, \quad y_2^* + y_3^* = 0. \tag{4.16}$$

Now let $z := (0, 1)$. Then we have $h'(\bar{x})z = 0$, implying $z \in \mathcal{C}(\bar{x})$, and for any sequence $\vec{t} \in \mathbb{T}$ the set $C''_{h'(\bar{x}), \vec{t}}(h(\bar{x}); z)$ is formed by limits of sequences $(y_0^n, y_1^n, y_2^n, y_3^n)$ such that

$$t_n z_1^n + \frac{t_n^2}{2} y_0^n \leq 0,$$
$$t_n z_1^n + \frac{t_n^2}{2} y_1^n \geq \left| \frac{t_n^2}{2} y_2^n \right|^{\frac{3}{2}},$$
$$t_n z_1^n + \frac{t_n^2}{2} y_3^n = t_n z_1^n + \frac{t_n^2}{2}(y_1^n + y_2^n)$$

hold for some sequence $(z_1^n, z_2^n) \to z$. It follows that

$$C''_{h'(\bar{x}), \vec{t}}(h(\bar{x}); z) = \{(y_0, y_1, y_2, y_3) : y_0 \leq y_1,\ y_3 = y_1 + y_2\}.$$

Now for $\alpha \in [0, 2\pi]$ define the sequence $\vec{t}^{\,\alpha} \in \mathbb{T}$ by $t_n^\alpha := e^{-2n\pi + \alpha}$ for all n. Then some straightforward calculations give

$$h''_{\vec{t}^{\,\alpha}}(\bar{x}; z) = \begin{pmatrix} -2 \\ -\frac{1}{2} \\ 0 \\ 1 + 2\sin\alpha \end{pmatrix}, \quad (h')'_{\vec{t}^{\,\alpha}}(\bar{x}; z) = \begin{pmatrix} 0 & -2 \\ 0 & -\frac{1}{2} \\ 1 & 0 \\ 0 & 1 + 2\sin\alpha + \cos\alpha \end{pmatrix},$$

and by taking $\alpha = -\arcsin\frac{3}{4}$ we obtain $h''_{\vec{t}^\alpha}(\bar{x};z) \in C''_{h'(\bar{x}),\vec{t}^\alpha}(h(\bar{x});z)$. Since Y is finite-dimensional, Assumption 2 holds, and hence we can apply Theorem 4.6. Assuming $\bar{x}$ to be a locally optimal solution, there exist multipliers $y^* \in \Lambda_{FJ} \cap \mathcal{S}_{Y^*}$ and $\mu^* \in \operatorname{Ker}(h'(\bar{x}))^\perp = \{(\mu_1,0) : \mu_1 \in \mathbb{R}\}$ such that conditions (4.3) and (4.4) with $s = z$ and $y = 0$ are fulfilled:

$$(h')'_{\vec{t}^\alpha}(\bar{x};z)^* y^* + \mu^* = \begin{pmatrix} y_2^* + \mu_1 \\ -2y_0^* - \frac{1}{2}y_1^* + y_3^*(1 + 2\sin\alpha + \cos\alpha) \end{pmatrix} = 0,$$
$$0 + \frac{1}{2}\langle y^*, h''_{\vec{t}^\alpha}(\bar{x};z) - 0\rangle = \frac{1}{2}\Big(-2y_0^* - \frac{1}{2}y_1^* + (1+2\sin\alpha)y_3^*\Big) \geq 0.$$

However, these conditions, together with condition (4.16), are fulfilled only for $y_0^* = y_1^* = y_2^* = y_3^* = \mu_1^* = 0$, contradicting $y^* \in \mathcal{S}_{Y^*}$. Hence, $\bar{x}$ is not a local minimizer.

5. Second-order necessary conditions for the scalar mathematical programming problem. We consider here the results of the preceding section for the special case of the scalar mathematical programming problem

$$\text{(MP)} \qquad \begin{aligned} &\min f(x) \\ &\text{s.t.} \quad g_i(x) \leq 0,\ i = 1,\dots,m, \\ &\qquad G(x) = 0, \end{aligned}$$

where $f : X \to \mathbb{R}$, $g_i : X \to \mathbb{R}$ for $i = 1,\dots,m$, $G : X \to \hat{V}$, and X and $\hat{V}$ are Banach spaces. Then $Y = \mathbb{R} \times \mathbb{R}^m \times \hat{V}$ and for a given feasible point $\bar{x}$ the mapping h and the set C according to (1.5) are given by $h(x) = (f(x) - f(\bar{x}), g_1(x),\dots,g_m(x), G(x))$ and $C = R_- \times \mathbb{R}^m_- \times \{0\}$. We will assume throughout this section that Assumption 1 holds. The set of multipliers Λ_{FJ} satisfying the first-order conditions of Fritz–John type consists of all multipliers $(\alpha,\lambda,v^*) \in \mathbb{R} \times \mathbb{R}^m \times \hat{V}^*$, such that

$$\begin{aligned} \mathcal{L}'_x(\bar{x},\alpha,\lambda,v^*) &= 0, \\ \alpha &\geq 0, \\ \lambda_i \geq 0,\ \lambda_i g_i(\bar{x}) &= 0,\ i = 1,\dots,m, \\ (\alpha,\lambda,v^*) &\neq (0,0,0), \end{aligned}$$

where the generalized Lagrangian $\mathcal{L}$ is given in the usual way by $\mathcal{L}(x,\alpha,\lambda,v^*) := \alpha f(x) + \sum_{i=1}^m \lambda_i g_i(x) + \langle v^*, G(x)\rangle$ and $\mathcal{L}'_x$ is the partial derivative of the Lagrangian with respect to x. For partial second-order directional derivatives of the Lagrangian with respect to the first variable we use the following notation:

$$\mathcal{L}''_{x\vec{t}}(\bar{x},\alpha,\lambda,v^*;z) := \alpha f''_{\vec{t}}(\bar{x};z) + \sum_{i=1}^m \lambda_i g_i{}''_{\vec{t}}(\bar{x};z) + \langle v^*, G''_{\vec{t}}(\bar{x};z)\rangle,$$
$$(\mathcal{L}'_x)'_{\vec{t}}(\bar{x},\alpha,\lambda,v^*;z) := \alpha(f')'_{\vec{t}}(\bar{x};z) + \sum_{i=1}^m \lambda_i (g_i')'_{\vec{t}}(\bar{x};z) + (G')'_{\vec{t}}(\bar{x};z)^* v^*.$$

Note that the multipliers (α,λ,v^*) form a linear functional $y^* \in Y^*$, and we have $\mathcal{L}'_x(\bar{x},\alpha,\lambda,v^*) = h'(\bar{x})^* y^*$, $\mathcal{L}''_{x\vec{t}}(\bar{x},\alpha,\lambda,v^*;z) = h''_{\vec{t}}(\bar{x};z)$ and $(\mathcal{L}'_x)'_{\vec{t}}(\bar{x},\alpha,\lambda,v^*;z) = (h')'_{\vec{t}}(\bar{x};z)^* y^*$.

In the case of the mathematical programming problem the second-order compound tangent sets have the property that

$$0 \in C''_{h'(\bar{x}),\vec{t}}(h(\bar{x});s) \quad \forall s \in \mathcal{C}(\bar{x}), \vec{t} \in \mathbb{T}. \tag{5.1}$$

This follows easily from the definition and the observation that for arbitrary $s \in \mathcal{C}(\bar{x})$ and $\vec{t} \in \mathbb{T}$ we have $h(\bar{x}) + t_n s \in C$ for all n sufficiently large. Condition (5.1) together with Lemma 4.2 implies

$$C''_{h'(\bar{x}),\vec{t}}(h(\bar{x}); s) = \mathrm{cl}\left(T_C(h(\bar{x})) + \mathrm{Im}\, h'(\bar{x})\right).$$

Hence, $C''_{h'(\bar{x}),\vec{t}}(h(\bar{x}); s)$ does not depend on the choice of the direction $s \in \mathcal{C}(\bar{x})$ and the sequence $\vec{t} \in \mathbb{T}$. Moreover, $C''_{h'(\bar{x}),\vec{t}}(h(\bar{x}); s)$ is a cone and its polar cone can be written as

$$C''_{h'(\bar{x}),\vec{t}}(h(\bar{x}); s)^\circ = N_C(h(\bar{x})) \cap (\mathrm{Im}\, h'(\bar{x}))^\perp = \Lambda_{FJ} \cup \{0\}.$$

It follows that $y \in C''_{h'(\bar{x}),\vec{t}}(h(\bar{x}); s)$ holds if and only if one has $\langle y^*, y\rangle \le 0$ for each $y^* \in \Lambda_{FJ}$.

In a next step, for fixed $z \in \mathcal{C}(\bar{x}) \cap \mathcal{S}_X$ and $\vec{t} \in \mathbb{T}$ with $h''_{\vec{t}}(\bar{x}; z) \in C''_{h'(\bar{x}),\vec{t}}(h(\bar{x}); z)$ we will analyze the conclusions (4.3) and (4.4) of Theorem 4.6 under condition (5.1). Using condition (4.4) with $s = z$ and $y = 0$ yields, together with the above characterization for $h''_{\vec{t}}(\bar{x}; z) \in C''_{h'(\bar{x}),\vec{t}}(h(\bar{x}); z)$,

$$\langle \bar{y}^*, h''_{\vec{t}}(\bar{x}; z)\rangle = \max_{y^* \in \Lambda_{FJ}} \langle y^*, h''_{\vec{t}}(\bar{x}; z)\rangle = 0.$$

Since $h''_{\vec{t}}(\bar{x}; z) \in C''_{h'(\bar{x}),\vec{t}}(h(\bar{x}); z)$ implies $h''_{\vec{t}}(\bar{x}; z) \in C''_{h'(\bar{x}),\vec{t}}(h(\bar{x}); s)$ for all $s \in \mathcal{C}(\bar{x})$ we obtain $\langle \bar{\mu}^*, z - s\rangle \ge 0$ for all $s \in \mathcal{C}(\bar{x})$. Thus $\bar{\mu}^* \in N_{\mathcal{C}(\bar{x})}(z)$ or, equivalently, since $\mathcal{C}(\bar{x})$ is a cone, $\bar{\mu}^* \in \{\mu^* \in \mathcal{C}(\bar{x})^\circ \,:\, \langle \mu^*, z\rangle = 0\}$, where $\mathcal{C}(\bar{x})^\circ$ denotes the polar cone of the critical cone $\mathcal{C}(\bar{x})$. Now in the case of (MP) the critical cone is given by

$$\mathcal{C}(\bar{x}) = \left\{ z \in X \,:\, \begin{array}{c} \langle f'(\bar{x}), z\rangle \le 0, \\ \langle g'_i(\bar{x}), z\rangle \le 0 \quad \forall i \in \bar{I}, \\ G'(\bar{x})z = 0 \end{array} \right\},$$

where $\bar{I} := \{i \in \{1, \dots, m\} \,:\, g_i(x) = 0\}$ denotes the index set of active inequality constraints. If $\mathrm{Im}\, G'(\bar{x})$ is closed, then by the generalized Farkas lemma (see, e.g., [10, Proposition 2.201]) the polar cone $\mathcal{C}(\bar{x})^\circ$ is given by the formula

$$\mathcal{C}(\bar{x})^\circ = \left\{ \alpha f'(x) + \sum_{i \in \bar{I}} \lambda_i g'_i(x)^* + G'(\bar{x})^* v^* \,:\, \alpha \ge 0, \lambda_i \ge 0, i \in \bar{I}, v^* \in \hat{V}^* \right\}.$$

Hence, $\bar{\mu}^* \in \{\mu^* \in \mathcal{C}(\bar{x})^\circ \,:\, \langle \mu^*, z\rangle = 0\}$ has the following representation:

$$\bar{\mu}^* = \tilde{\alpha} f'(x) + \sum_{i=1}^m \tilde{\lambda}_i g'_i(x) + G'(\bar{x})^* \tilde{v}^*,$$

where $\tilde{\alpha} \ge 0$, $\tilde{\lambda}_i \ge 0$, $\tilde{\lambda}_i g_i(\bar{x}) = 0$, $i = 1, \dots, m$, and $\tilde{v}^* \in \hat{V}^*$ are such that $\tilde{\alpha}\langle g'_i(\bar{x}), z\rangle + \sum_{i=0}^m \tilde{\lambda}_i \langle g'_i(\bar{x}), z\rangle + \langle \tilde{v}^*, G'(\bar{x})z\rangle = 0$, or, equivalently, since $z \in \mathcal{C}(\bar{x})$,

$$\tilde{\alpha}\langle f'(\bar{x}), z\rangle = 0, \quad \tilde{\lambda}_i g_i(\bar{x}) = \tilde{\lambda}_i \langle g'_i(\bar{x}), z\rangle = 0,\ i = 1, \dots, m.$$

Next let us consider Assumption 2. It surely holds if the space $\hat{V}^*$ is finite-dimensional. In the case of infinite-dimensional $\hat{V}$ let us examine the assumptions of

Theorem 4.8. Note that $\operatorname{aff}(C - h(\bar{x})) = \mathbb{R} \times \mathbb{R}^m \times \{0\}$ is always finite-dimensional for the problem (MP). From the discussion at the end of the preceding section we know that to verify 2-nondegeneracy of $h(\cdot)$ with respect to C we only have to consider the equality constraints $G(x) = 0$. It is straightforward to see that the mapping $G(\cdot)$ is 2-nondegenerate in direction z with respect to $\{0\}$ and the sequence $\vec{t} \in \mathbb{T}$, provided that $(G')'_{\vec{t}}(\bar{x}; z)$ exists if and only if $\operatorname{Im} G'(\bar{x})$ is closed in $\hat{V}$ and the mapping $s \to (G'(\bar{x})s, \pi G''_{\vec{t}}(\bar{x}; z)s)$ carries X onto $\operatorname{Im} G'(\bar{x}) \times \hat{V}/\operatorname{Im} G'(\bar{x})$, where π here denotes the quotient mapping onto the quotient space $\hat{V}/\operatorname{Im} G'(\bar{x})$. For twice, respectively, three times continuously differentiable mappings G this condition already appears in a lot of papers on optimality conditions for problems with degenerate equality constraints; see [4], [5], [25]. Further, in a slightly different version it is also known as the property of 2-*regularity* of the mapping G (see [6], [33]). We refer also to [19], where the theory of 2-regularity was applied to once differentiable mappings having a locally Lipschitzian derivative.

We summarize these considerations in the following corollary.

COROLLARY 5.1. *Let $\bar{x}$ be a local minimizer for the mathematical programming problem* (MP) *and let Assumption* 1 *hold. Then for each element $z \in \mathcal{C}(\bar{x})$ and each sequence $\vec{t} \in \mathbb{T}$ such that the second-order directional derivatives $h''_{\vec{t}}(\bar{x}; z)$ and $(h')'_{\vec{t}}(\bar{x}; z)$ exist, such that*

$$\sup_{(\alpha,\lambda,\hat{v}^*) \in \Lambda_{FJ}} \mathcal{L}''_{x\vec{t}}(\bar{x}, \alpha, \lambda, v^*; z) \le 0, \tag{5.2}$$

and such that either $\dim \hat{V} < \infty$ or $G(\cdot)$ is 2-nondegenerate in direction z with respect to $\{0\}$ and $(\vec{t})$, there exist multipliers $(\alpha, \lambda, v^*) \in \Lambda_{FJ}$ and $(\tilde{\alpha}, \tilde{\lambda}, \tilde{v}^*) \in \mathbb{R}_+ \times \mathbb{R}^m_+ \times \hat{V}^*$ such that

$$\mathcal{L}''_{x\vec{t}}(\bar{x}, \alpha, \lambda, v^*; z) = 0, \tag{5.3}$$

$$(\mathcal{L}'_x)'_{\vec{t}}(\bar{x}, \alpha, \lambda, v^*; z) + \mathcal{L}'_x(\bar{x}, \tilde{\alpha}, \tilde{\lambda}, \tilde{v}^*) = 0, \tag{5.4}$$

$$\tilde{\lambda}_i g_i(\bar{x}) = \tilde{\lambda}_i \langle g'_i(\bar{x}), z \rangle = 0, \ i = 1, \dots, m, \tag{5.5}$$

$$\tilde{\alpha} \langle f'(\bar{x}), z \rangle = 0. \tag{5.6}$$

Let us compare Corollary 5.1 with the standard second-order conditions (see, e.g., [16]) for (MP). For the sake of simplicity let us assume that f, g_i, $i = 1, \dots, m$, and G are twice continuously differentiable at $\bar{x}$ and that the range $\operatorname{Im} G'(\bar{x})$ is closed. If $\Lambda_{FJ} \neq \emptyset$ and if there is some $\beta > 0$ such that

$$\max_{(\alpha,\lambda,v^*) \in \Lambda_{FJ} \cap \mathcal{S}_{Y^*}} \mathcal{L}''_x(\bar{x}, \alpha, \lambda, v^*)(z, z) \ge \beta$$

for all $z \in \mathcal{C}(\bar{x}) \cap S_X$, then $\bar{x}$ is a strict local minimizer, and in fact one can show [15] that this condition is also equivalent for $\bar{x}$ to be an essential local minimizer of second order. On the other hand, the standard second-order necessary conditions state that at a local minimizer $\bar{x}$ the set Λ_{FJ} is not empty, and for each $z \in \mathcal{C}(\bar{x})$ there is some multiplier $(\alpha, \lambda, v^*) \in \Lambda_{FJ} \cap \mathcal{S}_{Y^*}$ such that

$$\mathcal{L}''_x(\bar{x}, \alpha, \lambda, v^*)(z, z) \ge 0.$$

It is also well known that on one hand this necessary condition is equivalent to condition (2.1) for nondegenerate points $\bar{x}$, i.e., when $\operatorname{Im} G'(\bar{x}) = \hat{V}$, and on the other hand that it is always satisfied when $\bar{x}$ is degenerate, whether $\bar{x}$ is a local minimizer or not.

Now, for degenerate points $\bar{x}$ Corollary 5.1 states the additional necessary conditions (5.4)–(5.6) for exactly those directions $z \in \mathcal{C}(\bar{x}) \cap \mathcal{S}_X$, where the sufficient conditions fail to hold, and thus reduces the gap between the standard necessary and sufficient conditions for the mathematical programming problem (MP).

Now let us compare our work with Avakov's results [4], [5]: For the sake of simplicity we consider only the case when no inequality constraints are present, i.e., $m = 0$. Avakov shows, assuming that f is Fréchet differentiable at $\bar{x}$ and G is twice Fréchet differentiable at a local minimizer $\bar{x}$ and the range $\operatorname{Im} G'(\bar{x})$ is closed, the following "first-order" conditions hold: For each $z \in X$ such that $G'(\bar{x})z = 0$, $G''(\bar{x})(z,z) \in \operatorname{Im} G'(\bar{x})$, and the operator $\mathcal{G}(\bar{x};z) : X \to \operatorname{Im} G'(\bar{x}) \times \hat{V}/\operatorname{Im} G'(\bar{x})$ given by $\mathcal{G}(\bar{x},z)s \to (G'(\bar{x})s, \pi G''(\bar{x})(z,s))$ has closed range, there exist multipliers $(0,0) \neq (\alpha, v^*) \in \mathbb{R}_+ \times \operatorname{Ker} G'(\bar{x})^*$ and $\tilde{v}^* \in \hat{V}^*$ such that

$$\alpha f'(\bar{x}) + G'(\bar{x})^*\tilde{v}^* + G''(\bar{x})(z,\cdot)^*v^* = 0. \tag{5.7}$$

Now let us demonstrate how this result follows from our work in the case in which the equality constraints are degenerate, i.e., $\operatorname{Im} G'(\bar{x}) \neq \hat{V}$, under the assumption that f is strictly differentiable at $\bar{x}$. Let $z \in X$ be fixed, satisfying $G'(\bar{x})z = 0$, $G''(\bar{x})(z,z) \in \operatorname{Im} G'(\bar{x})$. We can assume $\langle f'(\bar{x}), z\rangle \leq 0$, since otherwise we can take the direction $-z$. When $\mathcal{G}(\bar{x};z)$ is not surjective but has closed range, (5.7) with $\alpha = 0$ follows from the existence of a nontrivial functional in $(\operatorname{Im} \mathcal{G}(\bar{x};z))^\perp$ by using standard arguments.

Now assume that $\mathcal{G}(\bar{x};z)$ is surjective, i.e., G is 2-nondegenerate in direction z. It follows that $\pi G''(\bar{x})(z,\cdot) \neq 0$ and hence $z \neq 0$; w.l.o.g. $\|z\| = 1$. Let $p^* \in X^*$ denote a continuous linear functional with $\langle p^*, z\rangle = 1$ and consider the problem

$$(\mathrm{MP}_{\mathrm{p}^*}) \quad \min_{x \in X} \varphi(x) := \langle p^*, x - \bar{x}\rangle (f(x) - f(\bar{x})) \quad \text{s.t.} \quad \langle -p^*, x - \bar{x}\rangle \leq 0,\ G(x) = 0.$$

Then $\bar{x}$ is also a local minimizer for this problem and one can show that Assumption 1 is satisfied and also the second-order directional derivatives exist for every direction and every sequence $\vec{t} \in T$. In particular, we have $\varphi'(\bar{x}) = 0$, $\varphi''_{\vec{t}}(\bar{x};z) = 2\langle f'(\bar{x}), z\rangle\langle p, z\rangle$, and $(\varphi')'_{\vec{t}}(\bar{x};z) = \langle p^*, z\rangle f'(\bar{x}) + \langle f'(\bar{x}), z\rangle p^*$. It follows immediately that z belongs to the critical cone of problem $(\mathrm{MP}_{\mathrm{p}^*})$.

Now let $(0,0,0) \neq (\alpha, \lambda, v^*) \in \mathbb{R}_+ \times \mathbb{R}_+ \times \hat{V}^*$ be an arbitrary multiplier satisfying the Fritz–John conditions for $(\mathrm{MP}_{\mathrm{p}^*})$. We have $-\lambda p^* + G'(\bar{x})^*v^* = 0$ and consequently $\lambda = \lambda\langle p^*, z\rangle = \langle v^*, G'(\bar{x})z\rangle = 0$ and $G'(\bar{x})^*v^* = 0$. Due to our assumption on z we have $G''(\bar{x})(z,z) = G'(\bar{x})w$ for some $w \in X$ and therefore

$$\alpha\varphi''_{\vec{t}}(\bar{x};z) + 0 + \langle v^*, G''(\bar{x})(z,z)\rangle = 2\alpha\langle f'(\bar{x}), z\rangle\langle p^*, z\rangle + \langle v^*, G'(\bar{x})w\rangle = 2\alpha\langle f'(\bar{x}), z\rangle \leq 0.$$

This shows that condition (5.2) is satisfied and application of Corollary 5.1 proves the existence of multipliers $(0,0,0) \neq (\alpha, 0, v^*) \in \mathbb{R}_+ \times \mathbb{R}_+ \times \operatorname{Ker} G'(\bar{x})^*$ and $(\tilde{\alpha}, \tilde{\lambda}, \tilde{v}^*) \in \mathbb{R}_+ \times \mathbb{R}_+ \times \hat{V}^*$ such that we have $-\tilde{\lambda} = \tilde{\lambda}\langle -p^*, z\rangle = 0$ by condition (5.5), $\alpha\langle f'(\bar{x}), z\rangle = 0$ by condition (5.3), and also

$$\alpha(\varphi')'_{\vec{t}}(\bar{x};z) + G''(\bar{x})(z,\cdot)^*v^* + G'(\bar{x})^*\tilde{v}^* = \alpha f'(\bar{x}) + G'(\bar{x})^*\tilde{v}^* + G''(\bar{x})(z,\cdot)^*v^* = 0$$

by condition (5.4). Hence, condition (5.7) holds.

Note that condition (5.7) is called a "first-order" condition, although it contains the second derivative $G''(\bar{x})$. Similarly, Avakov and others presented "second-order" conditions (see [4], [5], [18], [25]), where third derivatives of the mapping G are involved. Of course, our results cannot cover such "second-order" conditions. Indeed,

as pointed out by an anonymous referee, the assertion of Corollary 5.1 follows from [5, Theorem 2] under such stronger differentiability assumptions.

There are also other second-order conditions known from the literature; see, for instance, the monograph of Arutyunov [2] and the references therein.

Example 5.1. This is a very easy example which can be treated by the results of this paper but not by results in the literature. Consider the problem

$$\min_{(x_1,x_2,x_3)\in\mathbb{R}^3} -x_1^2 + x_3 \quad \text{s.t.} \quad x_1x_2 + |x_1|^{5/2} = 0, \quad x_3 = 0,$$

at $\bar{x} = (0,0,0)$. Then $\bar{x} = (0,0,0)$ is not a local minimum, and this follows also from Corollary 5.1 since the second-order conditions (5.3)–(5.6) are not satisfied for the direction $z = (-1,0,0)$. However, the necessary "first-order" conditions (5.7) hold and the "second-order" conditions from [4], [25], [26] do not apply since the constraints are not three times differentiable at $\bar{x}$. Also the necessary conditions of Arutyunov [1, Theorems 3.1 and 3.2] and Belash and Tret'yakov [6, Theorem 3] are either satisfied or cannot be used since $f'(\bar{x}) \neq 0$ and $G'(\bar{x}) \neq 0$. Finally, if we replace the equality constraint $x_3 = 0$ by the inequality constraint $x_3 \geq 0$, then again the origin $\bar{x} = (0,0,0)$ can be classified as nonoptimal by Corollary 5.1, but the necessary conditions [3] cannot be used.

Example 5.2. Now consider the problem

$$\min_{(x_1,x_2)\in\mathbb{R}^2} x_1 \quad \text{s.t.} \quad x_2 \leq 0,\ x_1x_2 = 0,$$

at $\bar{x} = (0,0)$. Again, $\bar{x}$ is not a local minimum, but the second-order conditions of Corollary 5.1 now hold. To verify these conditions we have to consider the directions $z = (-1,0)$ and $z = (0,-1)$, and in both cases the multipliers $(\alpha,\lambda,v^*) \in \Lambda_{FJ}$ satisfying the conditions (5.3)–(5.6) are given by $(0,0,1)$. Moreover, for any direction $z \in \mathcal{C}(\bar{x}) \cap \mathcal{S}_X$ we have $\mathcal{L}''_{xt}(\bar{x},\alpha,\lambda,v^*;z) \geq 0$ for multipliers of the form $(\alpha,\lambda,v^*) = (0,0,t)$, $t > 0$.

Now we show that in a situation as in Example 5.2, where the multipliers corresponding to Corollary 5.1 are contained in a pointed closed convex cone, the second-order necessary conditions of Corollary 5.1 are sharp. We state this result in terms of the general problem (P).

THEOREM 5.2. *Let the point $\bar{x}$ be feasible for problem* (P) *and suppose that Assumption* 1 *holds, that* $\dim Y < \infty$, *and that*

$$\lim_{t\to 0_+} d\left(\frac{h(\bar{x}+tz) - h(\bar{x}) - th'(\bar{x})z}{t^2/2}, h''(\bar{x};z)\right) = 0 \tag{5.8}$$

holds uniformly for all $z \in \mathcal{C}(\bar{x}) \cap \mathcal{B}_X$. Further suppose that there is a pointed closed convex cone $\bar{\Lambda} \subset \Lambda_{FJ} \cup \{0\}$ such that for every $z \in \mathcal{C}(\bar{x})$ and every $y \in h''(\bar{x};z)$ there is some $\bar{y}^* \in \bar{\Lambda} \cap \mathcal{S}_{Y^*}$ with $\langle \bar{y}^*, y\rangle \geq 0$. Then there exists a mapping $\delta h = (\delta f, \delta g)$ with $\delta h(x) = \psi(\|x - \bar{x}\|)y$, where $y \in Y$ and $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ is a twice continuously differentiable function satisfying $\psi(0) = \psi'(0) = \psi''(0) = 0$, such that $\bar{x}$ is a strict local minimizer for (P) *with f and g replaced by $f + \delta f$ and $g + \delta g$, respectively.*

Proof. Let $S := \bar{\Lambda}^\circ$ denote the polar cone of the pointed convex closed cone $\bar{\Lambda}$. Since $\dim Y < \infty$, we have $\operatorname{int} S \neq \emptyset$. Further we have $\operatorname{cl}(h'(\bar{x})X + T_C(h(\bar{x}))) = \Lambda_{FJ}^\circ \subset S$. Let the subspace Q be given by $Q := \operatorname{aff}(h'(\bar{x})X - T_C(h(\bar{x})))$. If $\dim Q < \dim Y$, then we can find $p := \dim Y - \dim Q$ linearly independent elements $y_i \in S \setminus Q$, $i = 1,\dots,p$, forming a basis for some topological complement Q^c to Q, such that

$\operatorname{int}(h'(\bar{x})X - \bar{C}) \neq \emptyset$, where $\bar{C} := h(\bar{x}) + T_C(h(\bar{x})) + \hat{S}$ and the cone $\hat{S}$ is given by $\hat{S} := \{\sum_{i=1}^{p} \alpha_i y_i \,:\, \alpha_i \geq 0,\ i = 1, \dots, p\}$. Note that $\bar{C}$ is closed since $T_C(h(\bar{x})) \subset Q$ and $\hat{S} \subset Q^c$ are closed and $Y = Q \oplus Q^c$. On the other hand, if $\dim Q = \dim Y$, take $\bar{C} := h(\bar{x}) + T_C(h(\bar{x}) + \hat{S}$ with $\hat{S} = \{0\}$. In any case we have $\operatorname{int}(h'(\bar{x})X - \bar{C}) \neq \emptyset$ and $C \subset h(\bar{x}) + T_C(h(\bar{x})) \subset \bar{C} \subset h(\bar{x}) + S$. By taking into account the remark following Theorem 2.2 we see that in order to prove the theorem it is sufficient to show

$$\liminf_{\substack{x \to \bar{x} \\ \tau \to 0_+}} \frac{\hat{d}_{\bar{C}}(h(\bar{x}), h'(\bar{x}), \tau\|x - \bar{x}\|)}{\|x - \bar{x}\|^2} \geq 0.$$

Assume on the contrary that there are sequences $(z_n) \subset \mathcal{S}_X$, $\vec{t} = (t_n) \in \mathbb{T}$, and $(\tau_n) \in \mathbb{T}$, and a real $\beta > 0$ such that $t_n^{-2} \hat{d}_{\bar{C}}(h(\bar{x} + t_n z_n), h'(\bar{x}), \tau_n t_n) \leq -\beta$ for all n. Then by Theorem 2.4 we have $0 \in h(\bar{x} + t_n z_n) + \tau_n t_n h'(\bar{x})\mathcal{B}_X - \bar{C}$, and using Assumption 1 we see $t_n^{-1}(h(\bar{x} + t_n z_n) - h(\bar{x})) = h'(\bar{x})z_n + O(t_n) \in t_n^{-1}(\bar{C} - h(\bar{x})) = T_C(h(\bar{x})) + \hat{S}$. Since $\dim Y < \infty$, $h'(\bar{x})z_n \subset Y$ is a bounded sequence and $\operatorname{Im} h'(\bar{x})$ is closed, we have, by passing to a subsequence if necessary, $h'(\bar{x})z_n \to h'(\bar{x})\bar{z}$ for some $\bar{z} \in X$. Then $h'(\bar{x})\bar{z} \in T_C(h(\bar{x})) + \hat{S}$ and since $h'(\bar{x})\bar{z} \in Q$, $T_C(h(\bar{x}) \subset Q$, and $\hat{S} \cap Q = \{0\}$, $h'(\bar{x}) \in T_C(h(\bar{x}))$ follows, i.e., $\bar{z} \in \mathcal{C}(\bar{x})$. Further, since $\operatorname{Im} h'(\bar{x})$ is closed we can find another sequence, say (z_n'), such that $h'(\bar{x})z_n' = h'(\bar{x})\bar{z}$ and $\|z_n' - z_n\| \leq \gamma\|h'(\bar{x})\bar{z} - h'(\bar{x})z_n\|$ for some $\gamma > 0$. By taking $\tilde{z}_n := z_n'/\|z_n'\|$ we have found a sequence $(\tilde{z}_n) \subset \mathcal{C}(\bar{x}) \cap \mathcal{S}_X$ with $\tilde{z}_n - z_n \to 0$. By our assumptions, there exists a sequence (w_n) with $w_n \in h''(\bar{x}; \tilde{z}_n)$ such that $\|h(\bar{x} + t_n\tilde{z}_n) - (h(\bar{x}) + t_n h'(\bar{x})\tilde{z}_n + \frac{t_n^2}{2} w_n)\| = o(t_n^2)$ and, together with Assumption 1,

$$\left\| h(\bar{x} + t_n z_n) - \left(h(\bar{x}) + t_n h'(\bar{x}) z_n + \frac{t_n^2}{2} w_n \right) \right\| \leq \eta t_n^2 \|z_n - \tilde{z}_n\| + o(t_n^2) = o(t_n^2)$$

follows. Now, for each n let $y_n^* \in \bar{\Lambda} \cap \mathcal{S}_{Y^*}$ be chosen such that $\langle y_n^*, w_n \rangle \geq 0$. Since $y_n^* \in \Lambda_{FJ}$ we have $h'(\bar{x})^* y_n^* = 0$ and because of $h(\bar{x}) \in \bar{C}$ we obtain $\sigma_{\bar{C}}(y_n^*) \geq \langle y_n^*, h(\bar{x}) \rangle$. Hence we conclude that

$$\begin{aligned} t_n^{-2} \hat{d}_{\bar{C}}(h(\bar{x} + t_n z_n), h'(\bar{x}), \tau_n t_n) &\leq t_n^{-2}(\langle y_n^*, h(\bar{x} + t_n z_n) \rangle - \sigma_{\bar{C}}(y_n^*) - \tau_n t_n \|h'(\bar{x})^* y_n^*\|) \\ &\leq t_n^{-2} \langle y_n^*, h(\bar{x} + t_n z_n) - h(\bar{x}) \rangle \\ &= t_n^{-2} \langle y_n^*, h(\bar{x} + t_n z_n) - h(\bar{x}) - t_n h'(\bar{x}) z_n \rangle \\ &= \frac{1}{2} \langle y_n^*, w_n \rangle + o(1) \geq o(1), \end{aligned}$$

a contradiction. □

Note that, as a consequence of Assumption 1 and $\dim Y < \infty$, convergence in condition (5.8) is always uniform with respect to z in compact sets. Hence, besides the case when h is twice Fréchet differentiable at $\bar{x}$, condition (5.8) holds uniformly for all $z \in \mathcal{C}(\bar{x})$ when $\dim X < \infty$.

It is easy to see that the perturbation δh is Fréchet differentiable at $\bar{x}$ with $\delta h(\bar{x}) = 0$, $\delta h'(\bar{x}) = 0$ and that Assumption 1 is fulfilled with arbitrarily small η. Moreover, for any sequence $\vec{t} \in \mathbb{T}$ and any element $z \in X$ we have $h_{\vec{t}}''(\bar{x}; z) = 0$ and $(h')_{\vec{t}}'(\bar{x}; z) = 0$. Consequently, our assumptions are also fulfilled for the perturbed problem and all the quantities used in the optimality conditions do not change. Further note that δh is even twice continuously differentiable and $\delta h''(\bar{x}) = 0$, provided X is a Hilbert space.

In the case when f is scalar and K is a polyhedral cone, using the notion of 2-*normal mappings*, Arutyunov [1], [2] presented conditions which are sufficient for the

existence of a cone $\bar{\Lambda}$ satisfying the assumptions of Theorem 5.2. Further, Arutyunov showed 2-normal mappings to be generic under certain circumstances. However, note that the constraint mapping of Example 5.2 is not 2-normal.

Arutyunov [1, Theorem 4.3] stated also a result which is very similar to Theorem 5.2 and from which he concluded that the "gap" between his necessary and sufficient second-order conditions is as minimal as possible. Note that the second derivatives of Arutyunov's perturbations can be made only arbitrary small, but they do not vanish at the point $\bar{x}$ under consideration.

REFERENCES

[1] A. V. Arutyunov, *Second-order conditions in extremal problems. The abnormal points*, Trans. Amer. Math. Soc., 350 (1998), pp. 4341–4365.

[2] A. V. Arutyunov, *Optimality Conditions: Abnormal and Degenerate Problems*, Math. Appl. 526, Kluwer Academic, Dordrecht, Boston, London, 2000.

[3] A. V. Arutyunov and V. Jacimovic, *To the theory of extremum for abnormal problems*, Moscow Univ. Comput. Math. Cybernet., 1 (2000), pp. 30–37.

[4] E. R. Avakov, *Extremum conditions for smooth problems with equality-type constraints*, U.S.S.R. Comput. Math. and Math. Phys., 25 (1985), pp. 24–32.

[5] E. R. Avakov, *Necessary extremum conditions for smooth abnormal problems with equality and inequality-type constraints*, Math. Notes, 45 (1989), pp. 431–437.

[6] K. N. Belash and A. A. Tret'yakov, *Methods for solving degenerate problems*, U.S.S.R. Comput. Math. and Math. Phys., 28 (1988), pp. 90–94.

[7] A. Ben-Tal, *Second order and related extremality conditions in nonlinear programming*, J. Optim. Theory Appl., 31 (1980), pp. 143–165.

[8] A. Ben-Tal and J. Zowe, *A unified theory of first and second order conditions for extremum problems in topological vector spaces*, Math. Programming Stud., 19 (1982), pp. 39–76.

[9] J. F. Bonnans, R. Cominetti, and A. Shapiro, *Second order optimality conditions based on parabolic second order tangent sets*, SIAM J. Optim., 9 (1999), pp. 466–492.

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

[11] O. A. Brezhneva and A. A. Tret'yakov, *Optimality conditions for degenerate extremum problems with equality constraints*, SIAM J. Control Optim., 42 (2003), pp. 729–745.

[12] M. Buchner, J. E. Marsden, and S. Schecter, *Applications of the blowing-up construction and algebraic geometry to bifurcation problems*, J. Differential Equations, 48 (1983), pp. 404–433.

[13] A. Cambini, L. Martein, and R. Cambini, *A new approach to second-order optimality conditions in vector optimization*, in Advances in Multiple Objective and Goal Programming, Lecture Notes in Econom. and Math. Systems 455, R. Caballero, F. Ruiz, and R. E. Steuer, eds., Springer, Berlin, 1997, pp. 219–227.

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

[15] H. Gfrerer, *Second-order optimality conditions for scalar and vector optimization problems in Banach spaces*, SIAM J. Control Optim., 45 (2006), pp. 972–997.

[16] A. D. Ioffe, *Necessary and sufficient conditions for a local minimum.* 3: *Second order conditions and augmented duality*, SIAM J. Control Optim., 17 (1979), pp. 266–288.

[17] A. D. Ioffe, *On some recent developments in the theory of second order optimality conditions*, in Optimization. Proceedings of the Fifth French-German Conference (Varetz, 1988), S. Dolecki, ed., Lecture Notes in Math. 1405, Springer, Berlin, 1989, pp. 55–68.

[18] A. F. Izmailov and M. V. Solodov, *Optimality conditions for irregular inequality-constrained problems*, SIAM J. Control Optim., 40 (2001), pp. 1280–1295.

[19] A. F. Izmailov and M. V. Solodov, *The theory of* 2*-regularity for mappings with Lipschitzian derivatives and its applications to optimality conditions*, Math. Oper. Res., 27 (2002), pp. 614–635.

[20] B. Jiménez and V. Novo, *First- and second-order sufficient conditions for strict minimality in multiobjective programming*, Numer. Funct. Anal. Optim., 23 (2002), pp. 303–322.

[21] B. Jiménez and V. Novo, *Second-order necessary conditions in set constrained differentiable vector optimization*, Math. Methods Oper. Res., 58 (2003), pp. 299–317.

[22] B. Jiménez and V. Novo, *Optimality conditions in differentiable vector optimization via*

second-order tangent sets, Appl. Math. Optim., 49 (2004), pp. 123–144.
[23] H. Kawasaki, *An envelope-like effect of infinitely many inequality constraints on second-order necessary conditions for minimization problems*, Math. Program., 41 (1988), pp. 73–96.
[24] D. Klatte and B. Kummer, *Nonsmooth Equations in Optimization. Regularity, Calculus, Methods, and Applications*, Nonconvex Optim. Appl. 60, Kluwer Academic, Dordrecht, Boston, London, 2002.
[25] U. Ledzewicz and H. Schättler, *Second-order conditions for extremum problems with non-regular equality constraints*, J. Optim. Theory Appl., 86 (1995), pp. 113–144.
[26] U. Ledzewicz and H. Schättler, *High-order approximations and generalized necessary conditions for optimality*, SIAM J. Control Optim., 37 (1998), pp. 33–53.
[27] E. S. Levitin, A. A. Milyutin, and N. P. Osmolovskii, *Conditions of high order for a local minimum in problems with constraints*, Russian Math. Surveys, 33 (1978), pp. 97–168.
[28] J.-P. Penot, *Optimality conditions in mathematical programming and composite optimization*, Math. Program., 67 (1994), pp. 225–245.
[29] J.-P. Penot, *Second-order conditions for optimization problems with constraints*, SIAM J. Control Optim., 37 (1998), pp. 303–318.
[30] S. M. Robinson, *Stability theorems for systems of inequalities, part* II: *Differentiable nonlinear systems*, SIAM J. Numer. Anal., 13 (1976), pp. 497–513.
[31] S. M. Robinson, *Regularity and stability for convex multivalued functions*, Math. Oper. Res., 1 (1976), pp. 130–143.
[32] S. M. Robinson, *First order conditions for general nonlinear optimization*, SIAM J. Appl. Math., 30 (1976), pp. 597–607.
[33] A. A. Tret'yakov, *Necessary and sufficient conditions for optimality of p-th order*, U.S.S.R. Comput. Math. and Math. Phys., 24 (1984), pp. 123–127.

SIAM J. OPTIM.
Vol. 18, No. 2, pp. 613–642

MAJORIZING FUNCTIONS AND CONVERGENCE OF THE GAUSS–NEWTON METHOD FOR CONVEX COMPOSITE OPTIMIZATION*

CHONG LI† AND K. F. NG‡

Abstract. We introduce a notion of quasi regularity for points with respect to the inclusion $F(x) \in C$, where F is a nonlinear Fréchet differentiable function from $\mathbb{R}^v$ to $\mathbb{R}^m$. When C is the set of minimum points of a convex real-valued function h on $\mathbb{R}^m$ and F' satisfies the L-average Lipschitz condition of Wang, we use the majorizing function technique to establish the semilocal linear/quadratic convergence of sequences generated by the Gauss–Newton method (with quasi-regular initial points) for the convex composite function $h \circ F$. Results are new even when the initial point is regular and F' is Lipschitz.

Key words. the Gauss–Newton method, convex composite optimization, majorizing function, convergence

AMS subject classifications. Primary, 47J15, 65H10; Secondary, 41A29

DOI. 10.1137/06065622X

1. Introduction.

The convex composite optimization to be considered is as follows:

$$\min_{x \in \mathbb{R}^v} f(x) := h(F(x)), \tag{1.1}$$

where h is a real-valued convex function on $\mathbb{R}^m$ and F is a nonlinear Fréchet differentiable map from $\mathbb{R}^v$ to $\mathbb{R}^m$ (with norm $\|\cdot\|$). We assume throughout that h attains its minimum $h_{\min}$.

This problem has recently received a great deal of attention. As observed by Burke and Ferris in their seminal paper [4], a wide variety of its applications can be found throughout the mathematical programming literature, especially in convex inclusion, minimax problems, penalization methods, and goal programming (see also [2, 6, 7, 15, 22]). The study of (1.1) provides not only a unifying framework for the development and analysis of algorithms for solutions but also a convenient tool for the study of first- and second-order optimality conditions in constrained optimization [3, 5, 7, 22]. As in [4, 13], the study of (1.1) naturally relates to the convex inclusion problem

$$F(x) \in C, \tag{1.2}$$

where

$$C := \operatorname{argmin} h, \tag{1.3}$$

*Received by the editors April 4, 2006; accepted for publication (in revised form) January 16, 2007; published electronically August 1, 2007.

http://www.siam.org/journals/siopt/18-2/65622.html

†Department of Mathematics, Zhejiang University, Hangzhou 310027, P. R. China (cli@zju.edu.cn). This author was supported in part by the National Natural Science Foundation of China (grant 10671175) and the Program for New Century Excellent Talents in University.

‡Department of Mathematics, Chinese University of Hong Kong, Hong Kong, P. R. China (kfng@math.cuhk.edu.hk). This author was supported by a direct grant (CUHK) and an Earmarked Grant from the Research Grant Council of Hong Kong.

the set of all minimum points of h. Of course, it is meaningful to study (1.2) in its own right for a general closed convex set (cf. [14, 16]). In section 3, we introduce a new notion of quasi regularity for $x_0 \in \mathbb{R}^v$ with respect to the inclusion (1.2). This new notion covers the case of regularity studied by Burke and Ferris [4] as well as the case when $F'(x_0) - C$ is surjective, employed by Robinson [18]. More importantly, we introduce notions of the quasi-regular radius r_{x_0} and of the quasi-regular bound function β_{x_0} attached to each quasi-regular point x_0. For the general case this pair (r_{x_0}, β_{x_0}) together with a suitable Lipschitz-type assumption on F' enables us to address the issue of convergence of Gauss–Newton sequences provided by the following well-known algorithm (cf. [4, 10, 13, 31]).

ALGORITHM A (η, Δ, x_0). Let $\eta \in [1, +\infty)$, $\Delta \in (0, +\infty]$, and for each $x \in \mathbb{R}^v$ define $D_\Delta(x)$ by

$$D_\Delta(x) = \{d \in \mathbb{R}^v : \|d\| \le \Delta, h(F(x) + F'(x)d) \le h(F(x) + F'(x)d') \quad \forall d' \in \mathbb{R}^v \text{ with } \|d'\| \le \Delta\}. \tag{1.4}$$

Let $x_0 \in \mathbb{R}^v$ be given. For $k = 0, 1, \ldots$, having $x_0, x_1, \ldots, x_k$, determine x_{k+1} as follows.

If $0 \in D_\Delta(x_k)$, then stop; if $0 \notin D_\Delta(x_k)$, choose d_k such that $d_k \in D_\Delta(x_k)$ and

$$\|d_k\| \le \eta d(0, D_\Delta(x_k)), \tag{1.5}$$

and set $x_{k+1} = x_k + d_k$. Here $d(x, W)$ denotes the distance from x to W in the finite-dimensional Banach space containing W.

Note that $D_\Delta(x)$ is nonempty and is the solution set of the following convex optimization problem:

$$\min_{d \in \mathbb{R}^v, \|d\| \le \Delta} h(F(x) + F'(x)d), \tag{1.6}$$

which can be solved by standard methods such as the subgradient method, the cutting plane method, the bundle method, etc. (cf. [9]).

If the initial point x_0 is a quasi-regular point with (r_{x_0}, β_{x_0}) and if F' satisfies a Lipschitz-type condition (introduced by Wang [28]) with an absolutely continuous function L satisfying a suitable property in relation to (r_{x_0}, β_{x_0}), our main results presented in section 4 show that the Gauss–Newton sequence $\{x_n\}$ provided by Algorithm A (η, Δ, x_0) converges at a quadratic rate to some x^* with $F(x^*) \in C$ (in particular, x^* solves (1.1)). Even in the special case when x_0 is regular and F' is Lipschitz, the advantage of allowing β_{x_0} and L to be functions (rather than constants) provides results which are new even for the above special case. Examples are given in section 6 to show there are situations where our results are applicable but not the earlier results in the literature; in particular, Example 6.1 is a simple example to demonstrate a quasi-regular point which is not regular. We shall show that the Gauss–Newton sequence $\{x_n\}$ is "majorized" by the corresponding numerical sequence $\{t_n\}$ generated by the classical Newton method with initial point $t_0 = 0$ for a "majorizing" function of the following type (again introduced by Wang [28]):

$$\phi_\alpha(t) = \xi - t + \alpha \int_0^t L(u)(t - u)\, \mathrm{d}u \quad \text{for each } t \ge 0, \tag{1.7}$$

where ξ, α are positive constants and L is a positive-valued increasing (more precisely, nondecreasing) absolutely continuous function on $[0, +\infty)$. In the case when L is a

constant function, (1.7) reduces to

$$\phi_\alpha(t) = \xi - t + \frac{\alpha L}{2}t^2 \quad \text{for each } t \geq 0, \tag{1.8}$$

the majorizing function used by Kantorovich [11] and by Kantorovich and Akilov [12]. In the case when

$$L(u) = \frac{2\gamma}{(1-\gamma u)^3}, \tag{1.9}$$

expression (1.7) reduces to

$$\phi_\alpha(t) = \xi - t + \frac{\alpha\gamma t^2}{1-\gamma t}, \tag{1.10}$$

the majorizing function that Wang made use of in his work [28] on approximate zeros of Smale (cf. [26]). Motivated by this and as an application of our results in section 4, we provide a sufficient condition ensuring that a point $x_0 \in \mathbb{R}^v$ will be an "approximate solution" of (1.1) in the sense that the Gauss–Newton sequence $\{x_n\}$ generated by Algorithm A (η, Δ, x_0) converges to a solution of (1.1) and satisfies the condition

$$\|x_{n+1} - x_n\| \leq \left(\frac{1}{2}\right)^{2^{n-1}} \|x_n - x_{n-1}\| \quad \text{for each } n = 1, 2, \ldots \tag{1.11}$$

(the last condition was used by Smale [26] in his study of approximate zeros for Newton's method).

2. Preliminaries. Let $\mathbf{B}(x, r)$ stand for the open ball in $\mathbb{R}^v$ or $\mathbb{R}^m$ with center x and radius r, while the corresponding closed ball is denoted by $\overline{\mathbf{B}(x, r)}$. Let W be a closed convex subset of $\mathbb{R}^v$ or $\mathbb{R}^m$. The negative polar of W is denoted by $W^\ominus$ and defined by

$$W^\ominus = \{z : \langle z, w\rangle \leq 0 \text{ for each } w \in W\}.$$

Let L be a positive-valued increasing absolutely continuous function on $[0, +\infty)$, and let $\alpha > 0$. Let $r_\alpha > 0$ and $b_\alpha > 0$ such that

$$\alpha \int_0^{r_\alpha} L(u)\ \mathrm{d}u = 1 \quad \text{and} \quad b_\alpha = \alpha \int_0^{r_\alpha} L(u)u\ \mathrm{d}u \tag{2.1}$$

(thus $b_\alpha < r_\alpha$). Let $\xi \geq 0$, and define

$$\phi_\alpha(t) = \xi - t + \alpha \int_0^t L(u)(t-u)\ \mathrm{d}u \quad \text{for each } t \geq 0. \tag{2.2}$$

Thus

$$\phi'_\alpha(t) = -1 + \alpha \int_0^t L(u)\ \mathrm{d}u, \quad \phi''_\alpha(t) = \alpha L(t) \quad \text{for each } t \geq 0, \tag{2.3}$$

and $\phi'''_\alpha(t)$ exists almost everywhere thanks to the assumption that L is absolutely continuous. Let $t_{\alpha,n}$ denote the sequence generated by Newton's method for ϕ_α with initial point $t_{\alpha,0} = 0$:

$$t_{\alpha,n+1} = t_{\alpha,n} - \phi'_\alpha(t_{\alpha,n})^{-1}\phi_\alpha(t_{\alpha,n}) \quad \text{for each } n = 0, 1, \ldots . \tag{2.4}$$

In particular, by (2.2) and (2.3),

$$t_{\alpha,1} = \xi. \tag{2.5}$$

Below we list a series of useful lemmas for our purpose. They are either known or can be verified easily by elementary methods (such as by differential calculus). In particular, Lemma 2.3 and Lemma 2.1(i) are taken from [28], while Lemma 2.1(ii) and (iii) are well known. Here we shall give a proof of Lemma 2.4 as an illustration.

LEMMA 2.1. *Suppose that* $0 < \xi \le b_\alpha$. *Then* $b_\alpha < r_\alpha$ *and the following assertions hold:*

(i) ϕ_α *is strictly decreasing on* $[0, r_\alpha]$ *and strictly increasing on* $[r_\alpha, +\infty)$ *with*

$$\phi_\alpha(\xi) > 0, \quad \phi_\alpha(r_\alpha) = \xi - b_\alpha \le 0, \quad \phi_\alpha(+\infty) \ge \xi > 0. \tag{2.6}$$

Moreover, if $\xi < b_\alpha$, ϕ_α *has two zeros, denoted respectively by* r_α^* *and* r_α^{**}, *such that*

$$\xi < r_\alpha^* < \frac{r_\alpha}{b_\alpha}\xi < r_\alpha < r_\alpha^{**}, \tag{2.7}$$

and if $\xi = b_\alpha$, *then* ϕ_α *has a unique zero* r_α^* *in* $(\xi, +\infty)$ *(in fact,* $r_\alpha^* = r_\alpha$*).*

(ii) $\{t_{\alpha,n}\}$ *is strictly monotonically increasing and converges to* r_α^*.

(iii) *The convergence of* $\{t_{\alpha,n}\}$ *is of quadratic rate if* $\xi < b_\alpha$, *and linear if* $\xi = b_\alpha$.

LEMMA 2.2. *Let* r_α, b_α, *and* ϕ_α *be defined by* (2.1) *and* (2.2). *Let* $\alpha' > \alpha$ *with the corresponding* $\phi_{\alpha'}$. *Then the following assertions hold:*

(i) *The functions* $\alpha \mapsto r_\alpha$ *and* $\alpha \mapsto b_\alpha$ *are strictly decreasing on* $(0, +\infty)$.

(ii) $\phi_\alpha < \phi_{\alpha'}$ *on* $(0, +\infty)$.

(iii) *The function* $\alpha \mapsto r_\alpha^*$ *is strictly increasing on the interval* $I(\xi)$, *where* $I(\xi)$ *denotes the set of all* $\alpha > 0$ *such that* $\xi \le b_\alpha$.

LEMMA 2.3. *Let* $0 \le c < +\infty$. *Define*

$$\chi(t) = \frac{1}{t^2}\int_0^t L(c+u)(t-u)\ \mathrm{d}u \qquad \text{for each } 0 \le t < +\infty. \tag{2.8}$$

Then χ *is increasing on* $[0, +\infty)$.

LEMMA 2.4. *Define*

$$\omega_\alpha(t) = \phi'_\alpha(t)^{-1}\phi_\alpha(t), \qquad \text{for each } t \in [0, r_\alpha^*).$$

Suppose that $0 < \xi \le b_\alpha$. *Then* ω_α *is increasing on* $[0, r_\alpha^*)$.

Proof. Since

$$\omega'_\alpha(t) = \frac{\phi'_\alpha(t)^2 - \phi_\alpha(t)\phi''_\alpha(t)}{\phi'_\alpha(t)^2} \quad \text{for each } t \in [0, r_\alpha^*),$$

it suffices to show that

$$\zeta_\alpha(t) := \phi'_\alpha(t)^2 - \phi_\alpha(t)\phi''_\alpha(t) \ge 0 \quad \text{for each } t \in [0, r_\alpha^*).$$

Since $\zeta_\alpha(r_\alpha^*) = \phi'_\alpha(r_\alpha^*)^2 \ge 0$, it remains to show that ζ_α is decreasing on $[0, r_\alpha^*]$. To do this, note that by (2.3), ζ_α is absolutely continuous, and so the derivative of ζ_α exists almost everywhere on $[0, r_\alpha^*]$ with

$$\zeta'_\alpha(t) = \phi'_\alpha(t)\phi''_\alpha(t) - \phi_\alpha(t)\phi'''_\alpha(t) \le 0 \qquad \text{for a.e.} \quad t \in [0, r_\alpha^*),$$

because $\phi'_\alpha \le 0$ while $\phi_\alpha,\ \phi''_\alpha,\ \phi'''_\alpha \ge 0$ almost everywhere on $[0, r^*_\alpha)$. Therefore, ζ_α is decreasing on $[0, r^*_\alpha)$, and the proof is complete. ∎

The following conditions were introduced by Wang in [28] but using the terminologies of "the center Lipschitz condition with the L average" and "the center Lipschitz condition in the inscribed sphere with the L, average," respectively, for (a) and (b).

DEFINITION 2.5. *Let Y be a Banach space and let $x_0 \in \mathbb{R}^v$. Let G be a mapping from $\mathbb{R}^v$ to Y. Then G is said to satisfy*

(a) *the weak L-average Lipschitz condition on $\mathbf{B}(x_0, r)$ if*

$$\|G(x) - G(x_0)\| \le \int_0^{\|x-x_0\|} L(u)\,\mathrm{d}u \quad \text{for each } x \in \mathbf{B}(x_0, r); \tag{2.9}$$

(b) *the L-average Lipschitz condition on $\mathbf{B}(x_0, r)$ if*

$$\|G(x) - G(x')\| \le \int_{\|x'-x_0\|}^{\|x-x'\|+\|x'-x_0\|} L(u)\,\mathrm{d}u \quad \text{for all } x,\ x' \in \mathbf{B}(x_0, r) \tag{2.10}$$

$$\text{with } \|x - x'\| + \|x' - x_0\| \le r.$$

3. Regularities. Let C be a closed convex set in $\mathbb{R}^m$. Consider the inclusion

$$F(x) \in C. \tag{3.1}$$

Let $x \in \mathbb{R}^v$ and

$$\mathcal{D}(x) = \{d \in \mathbb{R}^v :\ F(x) + F'(x)d \in C\}. \tag{3.2}$$

Remark 3.1. In the case when C is the set of all minimum points of h and if there exists $d_0 \in \mathbb{R}^v$ with $\|d_0\| \le \Delta$ such that $d_0 \in \mathcal{D}(x)$, then $d_0 \in D_\Delta(x)$, and for each $d \in \mathbb{R}^v$ with $\|d\| \le \Delta$ one has

$$d \in D_\Delta(x) \Longleftrightarrow d \in \mathcal{D}(x) \Longleftrightarrow d \in D_\infty(x). \tag{3.3}$$

Remark 3.2. The set $\mathcal{D}(x)$ defined in (3.2) can be viewed as the solution set of the following "linearized" problem associated with (3.1):

$$(P_x): \qquad F(x) + F'(x)\,d \in C. \tag{3.4}$$

Thus $\beta(\|x - x_0\|)$ in (3.5) is an "error bound" in determining how far the origin is away from the solution set of (P_x).

DEFINITION 3.1. *A point $x_0 \in \mathbb{R}^v$ is called a quasi-regular point of the inclusion* (3.1) *if there exist $r \in (0, +\infty)$ and an increasing positive-valued function β on $[0, r)$ such that*

$$\mathcal{D}(x) \ne \emptyset \quad \text{and} \quad d(0, \mathcal{D}(x)) \le \beta(\|x - x_0\|)\, d(F(x), C) \quad \text{for all } x \in \mathrm{B}(x_0, r). \tag{3.5}$$

Let $\mathbf{r}_{x_0}$ denote the supremum of r such that (3.5) holds for some increasing positive-valued function β on $[0, r)$. Let $r \in [0, \mathbf{r}_{x_0}]$, and let $\mathcal{B}_r(x_0)$ denote the set of all increasing positive-valued functions β on $[0, r)$ such that (3.5) holds. Define

$$\beta_{x_0}(t) = \inf\{\beta(t) : \beta \in \mathcal{B}_{\mathbf{r}_{x_0}}(x_0)\} \quad \text{for each } t \in [0, \mathbf{r}_{x_0}). \tag{3.6}$$

Note that each $\beta \in \mathcal{B}_r(x_0)$ with $\lim_{t\to r^-} \beta(t) < +\infty$ can be extended to an element of $\mathcal{B}_{\mathbf{r}_{x_0}}(x_0)$. From this we can verify that

$$\beta_{x_0}(t) = \inf\{\beta(t) : \beta \in \mathcal{B}_r(x_0)\} \quad \text{for each } t \in [0, r). \tag{3.7}$$

We call $\mathbf{r}_{x_0}$ and β_{x_0} respectively the quasi-regular radius and the quasi-regular bound function of the quasi-regular point x_0.

DEFINITION 3.2. *A point $x_0 \in \mathbb{R}^v$ is a regular point of the inclusion* (3.1) *if*

$$\ker(F'(x_0)^T) \cap (C - F(x_0))^{\ominus} = \{0\}. \tag{3.8}$$

The notion of regularity relates to some other notions of regularity that can be found in the papers [1, 3, 20, 21, 25], which have played an important role in the study of nonsmooth optimizations. Some equivalent conditions on the regular points for (3.1) are given in [4]. In the following proposition the existence of constants r and β is due to Burke and Ferris [4], and the second assertion then follows from a remark after Definition 3.1.

PROPOSITION 3.3. *Let x_0 be a regular point of* (3.1). *Then there are constants $r > 0$ and $\beta > 0$ such that* (3.5) *holds for r and $\beta(\cdot) = \beta$; consequently, x_0 is a quasi-regular point with the quasi-regular radius $\mathbf{r}_{x_0} \geq r$ and the quasi-regular bound function $\beta_{x_0} \leq \beta$ on $[0, r]$.*

Another important link of the present study relates to Robinson's condition [18, 19] that the convex process $d \mapsto F'(x)d - C$ is onto $\mathbb{R}^m$. To see this, let us first recall the concept of convex process, which was introduced by Rockafeller [23, 24] for convexity problems (see also Robinson [19]).

DEFINITION 3.4. *A set-valued mapping $T : \mathbb{R}^v \to 2^{\mathbb{R}^m}$ is called a convex process from $\mathbb{R}^v$ to $\mathbb{R}^m$ if it satisfies*

(a) *$T(x + y) \supseteq Tx + Ty$ for all x, $y \in \mathbb{R}^v$;*

(b) *$T\lambda x = \lambda Tx$ for all $\lambda > 0$, $x \in \mathbb{R}^v$;*

(c) *$0 \in T0$.*

Thus $T : \mathbb{R}^v \to 2^{\mathbb{R}^m}$ is a convex process if and only if its graph $Gr(T)$ is a convex cone in $\mathbb{R}^v \times \mathbb{R}^m$. As usual, the domain, range, and inverse of a convex process T are respectively denoted by $\mathrm{D}(T)$, $\mathrm{R}(T)$, T^{-1}; i.e.,

$$\mathrm{D}(T) = \{x \in \mathbb{R}^v : Tx \neq \emptyset\},$$

$$\mathrm{R}(T) = \cup\{Tx : x \in \mathrm{D}(T)\},$$

$$T^{-1}y = \{x \in \mathbb{R}^v : y \in Tx\}.$$

Obviously T^{-1} is a convex process from $\mathbb{R}^m$ to $\mathbb{R}^v$. Furthermore, for a set A in an $\mathbb{R}^v$ or $\mathbb{R}^m$, it would be convenient to use the notation $\|A\|$ to denote its distance to the origin, that is,

$$\|A\| = \inf\{\|a\| : \ a \in A\}. \tag{3.9}$$

DEFINITION 3.5. *Suppose that T is a convex process. The norm of T is defined by*

$$\|T\| = \sup\{\|Tx\| : \ x \in \mathrm{D}(T), \ \ \|x\| \leq 1\}.$$

If $\|T\| < +\infty$, we say that the convex process T is normed.

For two convex processes T and S from $\mathbb{R}^v$ to $\mathbb{R}^m$, the addition and multiplication are defined respectively as follows:

$$(T+S)(x) = Tx + Sx \quad \text{for each } x \in \mathbb{R}^v,$$

$$(\lambda T)(x) = \lambda(Tx) \quad \text{for each } x \in \mathbb{R}^v \text{ and } \lambda \in \mathbb{R}.$$

Let C be a closed convex set in $\mathbb{R}^m$ and let $x \in \mathbb{R}^v$. We define T_x by

$$T_x d = F'(x)\, d - C \quad \text{for each } d \in \mathbb{R}^v. \tag{3.10}$$

Then its inverse is

$$T_x^{-1} y = \{d \in \mathbb{R}^v : \ F'(x)\, d \in y + C\} \quad \text{for each } y \in \mathbb{R}^m. \tag{3.11}$$

Note that T_x is a convex process in the case when C is a cone. Note also that $\mathrm{D}(T_x) = \mathbb{R}^v$ for each $x \in \mathbb{R}^v$ and $\mathrm{D}(T_{x_0}^{-1}) = \mathbb{R}^m$ if $x_0 \in \mathbb{R}^v$ is such that the following condition of Robinson is satisfied:

$$T_{x_0} \text{ carries } \mathbb{R}^v \text{ onto } \mathbb{R}^m. \tag{3.12}$$

Proposition 3.7 below shows that the condition of Robinson (3.12) implies that x_0 is a regular point of (3.1), and an estimate of the quasi-regular bound function is provided. For its proof we need the following lemma, which is known in [18].

LEMMA 3.6. *Let C be a closed convex cone in $\mathbb{R}^m$, and let $x_0 \in \mathbb{R}^v$ be such that the condition of Robinson* (3.12) *is satisfied. Then the following assertions hold:* (i) *$T_{x_0}^{-1}$ is normed.*

(ii) *If S is a linear transformation from $\mathbb{R}^v$ to $\mathbb{R}^m$ such that $\|T_{x_0}^{-1}\|\|S\| < 1$, then the convex process $\tilde{T}$ defined by*

$$\tilde{T} = T_{x_0} + S$$

carries $\mathbb{R}^v$ onto $\mathbb{R}^m$. Furthermore, $\tilde{T}^{-1}$ is normed and

$$\|\tilde{T}^{-1}\| \le \frac{\|T_{x_0}^{-1}\|}{1 - \|T_{x_0}^{-1}\|\|S\|}.$$

PROPOSITION 3.7. *Let $x_0 \in \mathbb{R}^v$, and let T_{x_0} be defined as in* (3.10). *Suppose that the condition of Robinson* (3.12) *is satisfied. Then the following assertions hold:*

(i) *x_0 is a regular point of* (3.1).

(ii) *Suppose further that C is a closed convex cone in $\mathbb{R}^m$ and that F' satisfies the weak L-average Lipschitz condition on $\mathbf{B}(x_0, r)$ for some $r > 0$. Let $\beta_0 = \|T_{x_0}^{-1}\|$ and let r_{β_0} be defined by*

$$\beta_0 \int_0^{r_{\beta_0}} L(u)\ \mathrm{d}u = 1 \tag{3.13}$$

(cf. (2.1)*). Then the quasi-regular radius $\mathbf{r}_{x_0}$ and the quasi-regular bound function β_{x_0} satisfy $\mathbf{r}_{x_0} \ge \min\{r, r_{\beta_0}\}$ and*

$$\beta_{x_0}(t) \le \frac{\beta_0}{1 - \beta_0 \int_0^t L(u)\ \mathrm{d}u} \quad \text{for each } t \text{ with } 0 \le t < \min\{r, r_{\beta_0}\}. \tag{3.14}$$

Proof. Suppose that the condition (3.12) is satisfied, and let y belong to the set of the intersection in (3.8). Then, in view of the definition of T_{x_0}, there exist $u \in \mathbb{R}^v$ and $c \in C$ such that $-y - F(x_0) = F'(x_0)u - c$. Hence, by (3.8),

$$\langle y, F'(x_0)u\rangle = \langle F'(x_0)^T y, u\rangle = 0 \quad \text{and} \quad \langle y, c - F(x_0)\rangle \le 0. \tag{3.15}$$

It follows that

$$\langle y, y\rangle = \langle y, c - F(x_0) - F'(x_0)u\rangle = \langle y, c - F(x_0)\rangle \le 0 \tag{3.16}$$

and hence $y = 0$. This shows that (3.8) holds, and so x_0 is a regular point of the inclusion (3.1).

Now let $r > 0$ and suppose that F' satisfies the weak L-average Lipschitz condition on $\mathbf{B}(x_0, r)$. Let $x \in \mathbb{R}^v$ such that $\|x - x_0\| < \min\{r, r_{\beta_0}\}$. Then

$$\|F'(x) - F'(x_0)\| \le \int_0^{\|x-x_0\|} L(u)\ \mathrm{d}u < \int_0^{r_{\beta_0}} L(u)\ \mathrm{d}u;$$

hence, by (3.13),

$$\|T_{x_0}^{-1}\|\|F'(x) - F'(x_0)\| < \|T_{x_0}^{-1}\| \int_0^{r_{\beta_0}} L(u)\ \mathrm{d}u = 1.$$

This with Lemma 3.6 implies that the convex process defined by

$$T_x d = F'(x)\, d - C = T_{x_0} d + [F'(x) - F'(x_0)]\, d \quad \text{for each } d \in \mathbb{R}^v$$

carries $\mathbb{R}^v$ onto $\mathbb{R}^m$ and

$$\|T_x^{-1}\| \le \frac{\|T_{x_0}^{-1}\|}{1 - \|T_{x_0}^{-1}\|\|F'(x) - F'(x_0)\|} \le \frac{\|T_{x_0}^{-1}\|}{1 - \|T_{x_0}^{-1}\| \int_0^{\|x-x_0\|} L(u)\ \mathrm{d}u}. \tag{3.17}$$

Since T_x is surjective, we have that $\mathcal{D}(x)$ is nonempty; in particular, for each $c \in C$,

$$T_x^{-1}(c - F(x)) \subseteq \mathcal{D}(x). \tag{3.18}$$

To see this, let $d \in T_x^{-1}(c - F(x))$. Then, by (3.11), one has that $F'(x)\, d \in c - F(x) + C \subseteq C - F(x)$, and so $F(x) + F'(x)\, d \in C$, that is, $d \in \mathcal{D}(x)$. Hence (3.18) is true. Consequently,

$$d(0, \mathcal{D}(x)) \le \|T_x^{-1}(c - F(x))\| \le \|T_x^{-1}\|\|c - F(x)\|.$$

Since this is valid for each $c \in C$, it is seen that

$$d(0, \mathcal{D}(x)) \le \|T_x^{-1}\| d(F(x), C).$$

Combining this with (3.17) and (3.6) gives the desired result (3.14), and the proof is complete. ∎

4. Convergence criterion. We assume throughout the remainder of this paper that C is the set of all minimum points of h. Let $x_0 \in \mathbb{R}^v$ be a quasi-regular point of the inclusion (3.1) with the quasi-regular radius $\mathbf{r}_{x_0}$ and the quasi-regular bound function β_{x_0}. Let $\eta \in [1, +\infty)$ and let

$$\xi := \eta \beta_{x_0}(0)\, d(F(x_0), C). \tag{4.1}$$

For all $\mathbf{r} \in (0, \mathbf{r}_{x_0}]$, we define

$$\alpha_0(\mathbf{r}) := \sup\left\{ \frac{\eta\beta_{x_0}(t)}{\eta\beta_{x_0}(t)\int_0^t L(u)\ \mathrm{d}u + 1} : \xi \le t < \mathbf{r} \right\}, \tag{4.2}$$

with the usual convention that $\sup\emptyset = -\infty$.

THEOREM 4.1. *Let $\eta \in [1,\infty)$ and $\Delta \in (0,\infty]$. Let $x_0 \in \mathbb{R}^v$ be a quasi-regular point of the inclusion* (3.1) *with the quasi-regular radius $\mathbf{r}_{x_0}$ and the quasi-regular bound function β_{x_0}. Let $\xi > 0$, $0 < \mathbf{r} \le \mathbf{r}_{x_0}$, and $\alpha_0(\mathbf{r})$ be as described above. Let $\alpha \ge \alpha_0(\mathbf{r})$ be a positive constant and let b_α, r_α be defined by* (2.1). *Let r_α^* denote the smaller zero of the function ϕ_α defined by* (2.2). *Suppose that F' satisfies the L-average Lipschitz condition on $\mathbf{B}(x_0, r_\alpha^*)$, and that*

$$\xi \le \min\{b_\alpha, \Delta\} \quad \textit{and} \quad r_\alpha^* \le \mathbf{r} \tag{4.3}$$

(for example, (4.3) *is satisfied if*

$$\xi \le \min\left\{ b_\alpha,\ \frac{b_\alpha}{r_\alpha}\mathbf{r},\ \Delta \right\} \tag{4.4}$$

holds). Let $\{x_n\}$ denote the sequence generated by Algorithm A (η, Δ, x_0). *Then, $\{x_n\}$ converges to some x^* such that $F(x^*) \in C$, and the following assertions hold for each $n = 1, 2, \ldots$:*

$$\|x_n - x_{n-1}\| \le t_{\alpha,n} - t_{\alpha,n-1}, \tag{4.5}$$

$$\|x_{n+1} - x_n\| \le (t_{\alpha,n+1} - t_{\alpha,n}) \left(\frac{\|x_n - x_{n-1}\|}{t_{\alpha,n} - t_{\alpha,n-1}} \right)^2, \tag{4.6}$$

$$F(x_n) + F'(x_n)(x_{n+1} - x_n) \in C, \tag{4.7}$$

and

$$\|x_{n-1} - x^*\| \le r_\alpha^* - t_{\alpha,n-1}. \tag{4.8}$$

Proof. By (2.7) and (4.4),

$$r_\alpha^* \le \frac{r_\alpha}{b_\alpha}\xi \le \mathbf{r}. \tag{4.9}$$

Hence (4.4)$\Longrightarrow$(4.3). Thus it suffices to prove the theorem for the case when (4.3) is assumed. By (4.3), (2.5), and Lemma 2.1, one has that, for each n,

$$\xi \le t_{\alpha,n} < r_\alpha^* \le \mathbf{r} \le \mathbf{r}_{x_0}. \tag{4.10}$$

By the quasi regularity assumption, it follows that

$$\mathcal{D}(x) \ne \emptyset \quad \text{and} \quad d(0, \mathcal{D}(x)) \le \beta_{x_0}(\|x - x_0\|)\, d(F(x), C) \quad \text{for each } x \in \mathbf{B}(x_0, \mathbf{r}). \tag{4.11}$$

Let $k \ge 1$. We use $\overline{1,k}$ to denote the set of all integers n satisfying $1 \le n \le k$. Below we will verify the following implication:

$$\begin{aligned} &\text{(4.5) holds for all } n \in \overline{1,k}, \text{ and (4.7) holds for } n = k-1 \\ &\qquad \Longrightarrow \text{(4.6) and (4.7) hold for } n = k. \end{aligned} \tag{4.12}$$

To do this, suppose that (4.5) holds for each $n \in \overline{1,k}$, and set

$$x_k^\tau = \tau x_k + (1-\tau)x_{k-1} \quad \text{for each } \tau \in [0,1]. \tag{4.13}$$

Note that

$$\|x_k - x_0\| \le \sum_{i=1}^{k} \|x_i - x_{i-1}\| \le \sum_{i=1}^{k} (t_{\alpha,i} - t_{\alpha,i-1}) = t_{\alpha,k} \tag{4.14}$$

and

$$\|x_{k-1} - x_0\| \le t_{\alpha,k-1} \le t_{\alpha,k}. \tag{4.15}$$

It follows from (4.13) and (4.10) that $x_k^\tau \in \mathbf{B}(x_0, r_\alpha^*) \subseteq \mathbf{B}(x_0, \mathbf{r})$ for each $\tau \in [0,1]$. Hence (4.11) holds for $x = x_k$, namely,

$$\mathcal{D}(x_k) \ne \emptyset \quad \text{and} \quad d(0, \mathcal{D}(x_k)) \le \beta_{x_0}(\|x_k - x_0\|)\, d(F(x_k), C). \tag{4.16}$$

We claim that

$$\eta d(0, \mathcal{D}(x_k)) \le (t_{\alpha,k+1} - t_{\alpha,k}) \left(\frac{\|x_k - x_{k-1}\|}{t_{\alpha,k} - t_{\alpha,k-1}} \right)^2 \le t_{\alpha,k+1} - t_{\alpha,k} \tag{4.17}$$

(the second inequality needs no proof by the assumption of (4.12)). To show the first inequality, using (4.7) for $n = k-1$ and the fact that F' satisfies an L-average Lipschitz condition on $\mathbf{B}(x_0, r_\alpha^*)$, together with the elementary identity

$$\int_0^1 \int_A^{A+\tau B} L(u)\ \mathrm{d}u\, \mathrm{d}\tau = \int_0^B L(A+u)\left(1 - \frac{u}{B}\right)\ \mathrm{d}u \quad \text{for all } A,\ B > 0, \tag{4.18}$$

we have by (4.16) that

$$\begin{aligned}
\eta d(0, \mathcal{D}(x_k)) &\le \eta \beta_{x_0}(\|x_k - x_0\|) d(F(x_k), C) \\
&\le \eta \beta_{x_0}(\|x_k - x_0\|) \|F(x_k) - F(x_{k-1}) - F'(x_{k-1})(x_k - x_{k-1})\| \\
&\le \eta \beta_{x_0}(\|x_k - x_0\|) \left\| \int_0^1 (F'(x_k^\tau) - F'(x_{k-1}))(x_k - x_{k-1})\ \mathrm{d}\tau \right\| \\
&\le \eta \beta_{x_0}(\|x_k - x_0\|) \int_0^1 \left(\int_{\|x_{k-1} - x_0\|}^{\tau\|x_k - x_{k-1}\| + \|x_{k-1} - x_0\|} L(u)\ \mathrm{d}u \right) \\
&\quad \times \|x_k - x_{k-1}\|\ \mathrm{d}\tau \\
&= \eta \beta_{x_0}(\|x_k - x_0\|) \\
&\quad \times \left(\int_0^{\|x_k - x_{k-1}\|} L(\|x_{k-1} - x_0\| + u)(\|x_k - x_{k-1}\| - u)\ \mathrm{d}u \right) \\
&\le \eta \beta_{x_0}(t_{\alpha,k}) \left(\int_0^{\|x_k - x_{k-1}\|} L(t_{\alpha,k-1} + u)(\|x_k - x_{k-1}\| - u)\ \mathrm{d}u \right),
\end{aligned}$$

where the last inequality is valid because L and β_{x_0} are increasing and thanks to (4.14) and (4.15). Since (4.5) holds for $n = k$, Lemma 2.3 implies that

$$\frac{\int_0^{\|x_k - x_{k-1}\|} L(t_{\alpha,k-1} + u)(\|x_k - x_{k-1}\| - u)\ \mathrm{d}u}{\|x_k - x_{k-1}\|^2}$$

$$\leq \frac{\int_0^{t_{\alpha,k} - t_{\alpha,k-1}} L(t_{\alpha,k-1} + u)(t_{\alpha,k} - t_{\alpha,k-1} - u)\ \mathrm{d}u}{(t_{\alpha,k} - t_{\alpha,k-1})^2},$$

and it follows from the earlier estimate that

$$\eta d(0, \mathcal{D}(x_k)) \leq \eta\beta_{x_0}(t_{\alpha,k})\left(\int_0^{t_{\alpha,k} - t_{\alpha,k-1}} L(t_{\alpha,k-1} + u)(t_{\alpha,k} - t_{\alpha,k-1} - u)\mathrm{d}u\right)$$

$$\times \left(\frac{\|x_k - x_{k-1}\|}{t_{\alpha,k} - t_{\alpha,k-1}}\right)^2. \tag{4.19}$$

Similarly by (2.2), (2.3), (2.4), and (4.18), we have

$$\begin{aligned}
\phi_\alpha(t_{\alpha,k}) &= \phi_\alpha(t_{\alpha,k}) - \phi_\alpha(t_{\alpha,k-1}) - \phi'_\alpha(t_{\alpha,k-1})(t_{\alpha,k} - t_{\alpha,k-1}) \\
&= \left(\int_0^1 [\phi'_\alpha(t_{\alpha,k-1} + \tau(t_{\alpha,k} - t_{\alpha,k-1})) - \phi'_\alpha(t_{\alpha,k-1})]\ \mathrm{d}\,\tau\right)(t_{\alpha,k} - t_{\alpha,k-1}) \\
&= \left(\alpha\int_0^1 \int_{t_{\alpha,k-1}}^{\tau(t_{\alpha,k} - t_{\alpha,k-1}) + t_{\alpha,k-1}} L(u)\ \mathrm{d}u\ \mathrm{d}\tau\right)(t_{\alpha,k} - t_{\alpha,k-1}) \\
&= \alpha\int_0^{t_{\alpha,k} - t_{\alpha,k-1}} L(t_{\alpha,k-1} + u)(t_{\alpha,k} - t_{\alpha,k-1} - u)\ \mathrm{d}u. \qquad (4.20)
\end{aligned}$$

On the other hand, by (4.10) and (4.2),

$$\frac{\eta\beta_{x_0}(t_{\alpha,k})}{\alpha_0(\mathbf{r})} \leq \left(1 - \alpha_0(\mathbf{r})\int_0^{t_{\alpha,k}} L(u)\,\mathrm{d}u\right)^{-1}.$$

Since $\alpha \geq \alpha_0(\mathbf{r})$ and by (2.3), it follows that

$$\frac{\eta\beta_{x_0}(t_{\alpha,k})}{\alpha} \leq \left(1 - \alpha\int_0^{t_{\alpha,k}} L(u)\ \mathrm{d}u\right)^{-1} = -(\phi'_\alpha(t_{\alpha,k}))^{-1}. \tag{4.21}$$

Combining (4.19)–(4.21) together with (2.4), the first inequality in (4.17) is seen to hold. Moreover, by Lemma 2.4 and (4.3), we have

$$t_{\alpha,k+1} - t_{\alpha,k} = -\phi'_\alpha(t_{\alpha,k})^{-1}\phi_\alpha(t_{\alpha,k}) \leq -\phi'_\alpha(t_{\alpha,0})^{-1}\phi_\alpha(t_{\alpha,0}) = \xi \leq \Delta,$$

so (4.17) implies that $d(0, \mathcal{D}(x_k)) \leq \Delta$. Hence there exists $d_0 \in \mathbb{R}^v$ with $\|d_0\| \leq \Delta$ such that $F(x_k) + F'(x_k)d_0 \in C$. Consequently, by Remark 3.1,

$$D_\Delta(x_k) = \{d \in \mathbb{R}^v : \|d\| \leq \Delta,\ F(x_k) + F'(x_k)d \in C\}$$

and

$$d(0, D_\Delta(x_k)) = d(0, \mathcal{D}(x_k)).$$

Since $d_k = x_{k+1} - x_k \in D_\Delta(x_k)$ by Algorithm A (η, Δ, x_0), it follows that (4.7) holds for $n = k$. Furthermore, one has that

$$\|x_{k+1} - x_k\| \le \eta d(0, D_\Delta(x_k)) = \eta d(0, \mathcal{D}(x_k)).$$

This with (4.17) yields that (4.6) holds for $n = k$, and hence implication (4.12) is proved.

Clearly, if (4.5) holds for each $n = 1, 2, \ldots$, then $\{x_n\}$ is a Cauchy sequence by the monotonicity of $\{t_n\}$ and hence converges to some x^*. Thus (4.8) is clear. Therefore, to prove the theorem, we need to prove only that (4.5), (4.6), and (4.7) hold for each $n = 1, 2, \ldots$. We will proceed by mathematical induction. First, by (4.1), (4.3), and (4.11), $\mathcal{D}(x_0) \ne \emptyset$ and

$$\eta d(0, \mathcal{D}(x_0)) \le \eta \beta_{x_0}(\|x_0 - x_0\|)\, d(F(x_0), C) = \eta \beta_{x_0}(0) d(F(x_0), C) = \xi \le \Delta.$$

Then using the same arguments just used above, we have that (4.7) holds for $n = 0$ and

$$\|x_1 - x_0\| = \|d_0\| \le \eta d(0, D_\Delta(x_0)) \le \eta \beta_{x_0}(0) d(F(x_0), C) = \xi = t_{\alpha,1} - t_{\alpha,0};$$

that is, (4.5) holds for $n = 1$. Thus, by (4.12), (4.6) and (4.7) hold for $n = 1$. Furthermore, assume that (4.5), (4.6), and (4.7) hold for all $1 \le n \le k$. Then

$$\|x_{k+1} - x_k\| \le (t_{\alpha,k+1} - t_{\alpha,k}) \left(\frac{\|x_k - x_{k-1}\|}{t_{\alpha,k} - t_{\alpha,k-1}} \right)^2 \le t_{\alpha,k+1} - t_{\alpha,k}.$$

This shows that (4.5) holds for $n = k + 1$, and hence (4.5) holds for all n with $1 \le n \le k+1$. Thus, (4.12) implies that (4.6) and (4.7) hold for $n = k+1$. Therefore, (4.5), (4.6), and (4.7) hold for each $n = 1, 2, \ldots$. The proof is complete. □

Remark 4.1. (a) In Theorem 4.1 if one assumes in addition that either (i) $\xi < b_\alpha$ or (ii) $\alpha > \alpha_0 := \alpha_0(\mathbf{r})$, then the convergence of the sequence $\{x_n\}$ is of quadratic rate. For the case of (i), this remark follows immediately from Lemma 2.1(iii) thanks to (4.5) and (4.8). If (ii) is assumed, then, by Lemma 2.2, $\xi \le b_\alpha < b_{\alpha_0}$ and $r^*_{\alpha_0} \le r^*_\alpha \le \mathbf{r}$. Hence (4.3) holds with α_0 in place of α. Since $\xi < b_{\alpha_0}$, we are now in the case (i) if α is replaced by α_0, and hence our remark here is established.

(b) Refinements for results presented in the remainder of this paper can also be established in a similar manner as in (a) above.

Remark 4.2. Suppose that there exists a pair $(\bar\alpha, \bar{\mathbf{r}})$ such that

$$\begin{cases} \bar\alpha = \alpha_0(\bar{\mathbf{r}}), \\ r^*_{\bar\alpha} = \bar{\mathbf{r}}. \end{cases} \tag{4.22}$$

Note that the function $\alpha \mapsto b_\alpha$ is decreasing by Lemma 2.2. Then, if (4.3) holds for some $(\alpha, \mathbf{r})$ with $\alpha \ge \bar\alpha$ and $r \ge \bar{\mathbf{r}}$, (4.3) does for $(\alpha, \mathbf{r}) = (\bar\alpha, \bar{\mathbf{r}})$ (and hence Theorem 4.1 is applicable).

Recall from Proposition 3.3 that the assumption for the existence of r, β in the following corollary is automatically satisfied when $x_0 \in \mathbb{R}^v$ is a regular point of the inclusion (3.1). This remark also applies to Theorems 5.1 and 5.6 and Corollary 5.2.

COROLLARY 4.2. *Let $x_0 \in \mathbb{R}^v$ be a regular point of the inclusion* (3.1) *with $r > 0$ and $\beta > 0$ such that*

$$\mathcal{D}(x) \ne \emptyset \quad \text{and} \quad d(0, \mathcal{D}(x)) \le \beta\, d(F(x), C) \quad \text{for all } x \in \mathrm{B}(x_0, r). \tag{4.23}$$

Let $\eta \in [1,\infty)$, $\Delta \in (0,\infty]$, $\xi = \eta\beta d(F(x_0), C)$,

$$\alpha = \frac{\eta\beta}{1+\eta\beta\int_0^{\xi} L(u)\mathrm{d}\,u}, \tag{4.24}$$

and let b_α, r_α be defined by (2.1). *Let r_α^* denote the smaller zero of the function ϕ_α defined by* (2.2). *Suppose that F' satisfies the L-average Lipschitz condition on $\mathbf{B}(x_0, r_\alpha^*)$ and that*

$$\xi \le \min\{b_\alpha, \Delta\} \quad \textit{and} \quad r_\alpha^* \le r \tag{4.25}$$

(for example, (4.25) *is satisfied if*

$$\xi \le \min\left\{b_\alpha,\ \frac{b_\alpha}{r_\alpha} r,\ \Delta\right\} \tag{4.26}$$

holds). Then the conclusions of Theorem 4.1 *hold.*

Proof. Note that $\xi < {r_\alpha}^* \le r$ by Lemma 2.1 and (4.25). By (3.6) and (3.7), it is clear that $\mathbf{r}_{x_0} \ge r$ and $\beta_{x_0}(\cdot) \le \beta$ on $[0, r)$. Let $\mathbf{r} := r$, and let $\alpha_0(\mathbf{r})$ be defined by (4.2) as in Theorem 4.1. Then, by (4.24), we have that

$$\alpha \ge \frac{\eta\beta_{x_0}(t)}{1+\eta\beta_{x_0}(t)\int_0^t L(u)\,\mathrm{d}u} \quad \text{for each } t \in [\xi, r).$$

Hence $\alpha \ge \alpha_0(\mathbf{r})$ by (4.2). Note that (4.26) (resp., (4.25)) is identical to (4.4) (resp., (4.3)); Theorem 4.1 is applicable, and the proof is complete. □

COROLLARY 4.3. *Let $\eta \in [1, +\infty)$, $\Delta \in (0, +\infty]$, and let C be a cone. Let $x_0 \in \mathbb{R}^v$ be such that T_{x_0} carries $\mathbb{R}^v$ onto $\mathbb{R}^m$. Let*

$$\xi = \eta\|T_{x_0}^{-1}\|\, d(F(x_0), C), \tag{4.27}$$

$$\alpha = \frac{\eta\|T_{x_0}^{-1}\|}{1+(\eta-1)\|T_{x_0}^{-1}\|\int_0^{\xi} L(u)\ \mathrm{d}u}, \tag{4.28}$$

and let b_α, r_α be defined by (2.1). *Let r_α^* denote the smaller zero of the function ϕ_α defined by* (2.2). *Suppose that F' satisfies the L-average Lipschitz condition on $\mathbf{B}(x_0, r_\alpha^*)$ and that*

$$\xi \le \min\{b_\alpha, \Delta\}. \tag{4.29}$$

Then the conclusions of Theorem 4.1 *hold.*

Proof. Let $\beta_0 = \|T_{x_0}^{-1}\|$, and let r_{β_0} be defined by (3.13). Then, by Proposition 3.7(ii), we know that x_0 is a quasi-regular point with the quasi-regular radius

$$\mathbf{r}_{x_0} \ge \min\{r_\alpha^*, r_{\beta_0}\} \tag{4.30}$$

and the quasi-regular bound function

$$\beta_{x_0}(t) \le \frac{\beta_0}{1-\beta_0\int_0^t L(u)\ \mathrm{d}u} \quad \text{for each } t \text{ with } 0 \le t < \min\{r_\alpha^*, r_{\beta_0}\}. \tag{4.31}$$

Let $\mathbf{r} := \min\{r^*_\alpha, r_{\beta_0}\}$, and let $\alpha_0(\mathbf{r})$ be defined by (4.2). We claim that

$$\alpha \geq \alpha_0(\mathbf{r}) \tag{4.32}$$

and

$$r_{\beta_0} \geq r^*_\alpha. \tag{4.33}$$

Granting this, the minimum on the right-hand side of (4.30) is simply r^*_α, and so $\mathbf{r} = r^*_\alpha \leq \mathbf{r}_{x_0}$. Moreover, we note that $\beta_{x_0}(0) \leq \beta_0$ by (4.31), and so the ξ defined by (4.1) is majorized by that defined by (4.27); thus (4.29) entails that (4.3) holds and Theorem 4.1 is applicable. Therefore we need only to prove our claim. Note by (4.31) that, for each $\xi \leq t < \min\{r^*_\alpha, r_{\beta_0}\} = \mathbf{r}$, we have

$$\eta \int_0^t L(u)\ \mathrm{d}u + \frac{1}{\beta_{x_0}(t)} \geq \frac{1}{\beta_0} + (\eta-1)\int_0^t L(u)\ \mathrm{d}u \geq \frac{1}{\beta_0} + (\eta-1)\int_0^\xi L(u)\ \mathrm{d}u; \tag{4.34}$$

that is,

$$\frac{\beta_{x_0}(t)}{1+\eta\beta_{x_0}(t)\int_0^t L(u)\mathrm{d}u} \leq \frac{\beta_0}{1+(\eta-1)\beta_0\int_0^t L(u)\mathrm{d}u}. \tag{4.35}$$

Thus (4.32) follows by definitions of $\alpha_0(\mathbf{r})$ and of α, respectively given by (4.2) and (4.28). To verify (4.33), consider the two cases (i) $\alpha \geq \beta_0$ and (ii) $\alpha < \beta_0$. In (i), since by Lemma 2.2, r_α is decreasing with respect to α, we have that $r^*_\alpha \leq r_\alpha \leq r_{\beta_0}$. In (ii), since r^*_α is increasing with respect to α by Lemma 2.2, we have that $r^*_\alpha \leq r^*_{\beta_0} \leq r_{\beta_0}$. Therefore, (4.33) holds in all cases, and the proof is complete. □

Remark 4.3. (a) If the strict inequalities in (4.25) (resp., (4.29)) of Corollary 4.2 (resp., Corollary 4.3) hold, then the starting point x_0 of the sequence $\{x_n\}$ can be replaced by a nearby point; that is, there exists a neighborhood $U(x_0)$ of x_0 such that the sequence $\{x_n\}$ generated by Algorithm A $(\eta, \Delta, \bar{x})$ with initial point $\bar{x}$ from $U(x_0)$ converges to some solution of the inclusion problem (3.1) at a quadratic rate.

(b) Refinements for results presented in the remainder of this paper can also be established in a similar manner as in (a) above.

5. Special cases and applications. This section is devoted to some applications. First we specialize results of the preceding section to two important cases of the function L: L =constant and $L = \frac{2\gamma}{(1-\gamma u)^3}$. Second, mimicking Smale's γ-theory about the approximation zeros for Newton's method in solving nonlinear equations, we do the same for the Gauss–Newton method in solving composite convex optimization.

5.1. Kantorovich type. Throughout this subsection, we assume that the function L is a constant function. Then, by (2.1) and (2.2), we have that, for all $\alpha > 0$,

$$r_\alpha = \frac{1}{\alpha L}, \quad b_\alpha = \frac{1}{2\alpha L} \tag{5.1}$$

and

$$\phi_\alpha(t) = \xi - t + \frac{\alpha L}{2}t^2.$$

Moreover, if $\xi \leq \frac{1}{2\alpha L}$, then the zeros of ϕ_α are given by

$$\left.\begin{matrix} r^*_\alpha \\ r^{**}_\alpha \end{matrix}\right\} = \frac{1 \mp \sqrt{1-2\alpha L\xi}}{\alpha L}. \tag{5.2}$$

It is also known (see, for example, [8, 17, 29]) that $\{t_{\alpha,n}\}$ has the closed form

$$t_{\alpha,n} = \frac{1 - q_\alpha^{2^n-1}}{1 - q_\alpha^{2^n}} r_\alpha^* \quad \text{for each } n = 0, 1, \ldots, \tag{5.3}$$

where

$$q_\alpha := \frac{r_\alpha^*}{r_\alpha^{**}} = \frac{1 - \sqrt{1 - 2\alpha L\xi}}{1 + \sqrt{1 - 2\alpha L\xi}}. \tag{5.4}$$

For the present case (L is a positive constant), a commonly used version of Lipschitz continuity on $\mathbf{B}(x_0, r)$ is of course the following: a function G is Lipschitz continuous with modulus L (Lipschitz constant) if

$$\|G(x_1) - G(x_2)\| \le L\|x_1 - x_2\| \quad \text{for all } x_1,\, x_2 \in \mathbf{B}(x_0, r).$$

Clearly, this is a stronger requirement than the corresponding ones given in Definition 2.5. Although the weaker requirement of Definition 2.5(b) is sufficient for results in this subsection, we prefer to use the Lipschitz continuity in this regard to be in line with the common practice.

THEOREM 5.1. *Let $x_0 \in \mathbb{R}^v$ be a regular point of the inclusion* (3.1) *with $r > 0$ and $\beta > 0$ such that* (4.23) *holds. Let $L \in (0, +\infty)$, $\eta \in [1, +\infty)$, $\Delta \in (0, +\infty]$, $\xi = \eta\beta d(F(x_0), C)$,*

$$R^* = \frac{1 + L\eta\beta\xi - \sqrt{1 - (L\eta\beta\xi)^2}}{L\eta\beta} \quad \textit{and} \quad Q = \frac{1 - \sqrt{1 - (L\eta\beta\xi)^2}}{L\eta\beta\xi}. \tag{5.5}$$

Assume that F' is Lipschitz continuous on $\mathbf{B}(x_0, R^*)$ with modulus L, and that

$$\xi \le \min\left\{\frac{1}{L\beta\eta},\ \Delta\right\} \quad \textit{and} \quad r \ge R^* \tag{5.6}$$

(for example, (5.6) *is satisfied if*

$$\xi \le \min\left\{\frac{1}{L\beta\eta},\ \frac{1}{2}r,\ \Delta\right\} \tag{5.7}$$

holds). Let $\{x_n\}$ denote the sequence generated by Algorithm A *(η, Δ, x_0). Then $\{x_n\}$ converges to some x^* with $F(x^*) \in C$ and*

$$\|x_n - x^*\| \le \frac{Q^{2^n-1}}{\sum_{i=0}^{2^n-1} Q^i} R^* \quad \textit{for each } n = 0, 1, \ldots. \tag{5.8}$$

Proof. Let α be given as in (4.24), namely, $\alpha = \frac{\eta\beta}{1+L\eta\beta\xi}$. Moreover, by (5.1)–(5.5), one has that

$$r_\alpha^* = R^*, \quad q_\alpha = Q, \quad r_\alpha = \frac{1 + L\eta\beta\xi}{L\eta\beta}, \quad b_\alpha = \frac{1 + L\eta\beta\xi}{2L\eta\beta} \tag{5.9}$$

and

$$t_{\alpha,n} = \frac{1 - Q^{2^n-1}}{1 - Q^{2^n}} R^*. \tag{5.10}$$

Hence condition (4.3) (resp., (4.4)) is equivalent to the three inequalities

$$\xi \le \frac{1+L\eta\beta\xi}{2L\eta\beta}, \quad \xi \le \Delta, \quad \text{and} \quad r^*_\alpha \le r \ \left(\text{resp., } \ \xi \le \frac{b_\alpha}{r_\alpha} r\right),$$

and is hence, by (5.9), also equivalent to condition (5.6) (resp., (5.7)). Thus we apply Corollary 4.2 to conclude that the sequence $\{x_n\}$ converges to some x^* with $F(x^*) \in C$ and, for each $n = 1, 2, \ldots,$

$$\|x_n - x^*\| \le r^*_\alpha - t_{\alpha,n}.$$

Noting, by (5.9) and (5.10), that

$$r^*_\alpha - t_{\alpha,n} = \left(1 - \frac{1-Q^{2^n-1}}{1-Q^{2^n}}\right) R^* = \frac{Q^{2^n-1}}{\sum_{i=0}^{2^n-1} Q^i} R^*,$$

it follows that (5.8) holds, and the proof is complete. □

The following corollary (which requires no proof by virtue of Theorem 5.1 and Remark 4.3(a)) is a slight extension of [13, Theorem 1] (which, in turn, extends a result of Burke and Ferris [4, Theorem 4.1]), and our conditions such as (5.11) are more direct than the corresponding ones in [13]. In fact, the conditions (a)–(c) of [13, Theorem 1] clearly imply the condition (5.11) below. Moreover, by (a) and (b) of [13, Theorem 1], $\bar\delta > 4\eta\bar\beta d(F(\bar x), C) = 4\bar\xi$. Since, for each $\xi \le \frac{1}{L\bar\beta\eta}$,

$$\frac{1+L\eta\bar\beta\bar\xi - \sqrt{1-(L\eta\bar\beta\bar\xi)^2}}{L\eta\bar\beta\bar\xi} \le 2,$$

one has that

$$\frac{1+L\eta\bar\beta\bar\xi - \sqrt{1-(L\eta\bar\beta\bar\xi)^2}}{L\eta\bar\beta} = \frac{1+L\eta\bar\beta\bar\xi - \sqrt{1-(L\eta\bar\beta\bar\xi)^2}}{L\eta\bar\beta\bar\xi}\bar\xi \le 2\bar\xi < \bar\delta.$$

Hence, $\bar r^* := \bar\delta$ satisfies the requirements of Corollary 5.2 below.

COROLLARY 5.2. *Let $\bar x \in \mathbb{R}^v$ be a regular point of the inclusion* (3.1) *with positive constants $\bar r$ and $\bar\beta$ satisfying* (4.23) *in place of r and β, respectively. Let $L \in (0, +\infty)$, $\eta \in [1, +\infty)$, $\Delta \in (0, +\infty]$, $\bar\xi = \eta\bar\beta d(F(\bar x), C)$, and let $\bar r^* > \frac{1+L\eta\bar\beta\bar\xi - \sqrt{1-(L\eta\bar\beta\bar\xi)^2}}{L\eta\bar\beta}$. Assume that F' is Lipschitz continuous on $\mathbf{B}(\bar x, \bar r^*)$ with modulus L, and that*

$$\bar\xi < \min\left\{\frac{1}{L\bar\beta\eta}, \ \frac{1}{2}\bar r, \ \Delta\right\}. \tag{5.11}$$

Then, there exists a neighborhood $U(\bar x)$ of $\bar x$ such that the sequence $\{x_n\}$ generated by Algorithm A *(η, Δ, x_0) with $x_0 \in U(\bar x)$ converges at a quadratic rate to some x^* with $F(x^*) \in C$, and the estimate* (5.8) *holds.*

THEOREM 5.3. *Let $\eta \in [1, +\infty)$, $\Delta \in (0, +\infty]$, and let C be a cone. Let $x_0 \in \mathbb{R}^v$ be such that T_{x_0} carries $\mathbb{R}^v$ onto $\mathbb{R}^m$. Let $L \in (0, +\infty)$ and $\xi = \eta\|T^{-1}_{x_0}\| d(F(x_0), C)$. Instead of* (5.5)*, we write*

$$R^* = \frac{1+(\eta-1)L\|T^{-1}_{x_0}\|\xi - \sqrt{1-2L\|T^{-1}_{x_0}\|\xi - (\eta^2-1)(L\|T^{-1}_{x_0}\|\xi)^2}}{L\|T^{-1}_{x_0}\|\eta} \tag{5.12}$$

and

$$Q = \frac{1 - L\|T_{x_0}^{-1}\|\xi - \sqrt{1 - 2L\|T_{x_0}^{-1}\|\xi - (\eta^2 - 1)(L\|T_{x_0}^{-1}\|\xi)^2}}{L\|T_{x_0}^{-1}\|\eta\xi}. \tag{5.13}$$

Suppose that F' is Lipschitz continuous on $\mathbf{B}(x_0, R^*)$ with modulus L, and that

$$\xi \le \min\left\{\frac{1}{L\|T_{x_0}^{-1}\|(\eta+1)}, \Delta\right\}. \tag{5.14}$$

Then, the same conclusions hold as in Theorem 5.1.

Proof. Let α be defined as in (4.28); that is, $\alpha = \frac{\eta\|T_{x_0}^{-1}\|}{1+(\eta-1)L\|T_{x_0}^{-1}\|\xi}$. Then, by (5.1), the following equivalences hold:

$$\xi \le b_\alpha \Longleftrightarrow 2L\xi\eta\|T_{x_0}^{-1}\| \le 1 + (\eta-1)L\|T_{x_0}^{-1}\|\xi \Longleftrightarrow \xi L\|T_{x_0}^{-1}\|(1+\eta) \le 1;$$

that is, (4.29) and (5.14) are equivalent. Moreover, it is easy to verify that R^* and Q defined in (5.12) and (5.13), respectively, are equal to r_α^* and q_α defined in (5.2) and (5.4). Therefore, one can complete the proof in the same way as for Theorem 5.1 but using Corollary 4.3 in place of Corollary 4.2. ∎

5.2. Smale's type. Let $\gamma > 0$. For the remainder of this section we assume that L is the function defined by

$$L(u) = \frac{2\gamma}{(1-\gamma u)^3} \quad \text{for each } u \text{ with } 0 \le u < \frac{1}{\gamma}. \tag{5.15}$$

Then, by (2.1), (2.2), and elementary calculation (cf. [28]), one has that for all $\alpha > 0$,

$$r_\alpha = \left(1 - \sqrt{\frac{\alpha}{1+\alpha}}\right)\frac{1}{\gamma}, \quad b_\alpha = \left(1 + 2\alpha - 2\sqrt{\alpha(1+\alpha)}\right)\frac{1}{\gamma} \tag{5.16}$$

and

$$\phi_\alpha(t) = \xi - t + \frac{\alpha\gamma t^2}{1-\gamma t} \quad \text{for each } t \text{ with } 0 \le t < \frac{1}{\gamma}. \tag{5.17}$$

Thus, from [28], we have the following lemma.

LEMMA 5.4. *Let $\alpha > 0$. Assume that $\xi \le b_\alpha$, namely,*

$$\gamma\xi \le 1 + 2\alpha - 2\sqrt{\alpha(1+\alpha)}. \tag{5.18}$$

Then the following assertions hold:

(i) *ϕ_α has two zeros given by*

$$\left.\begin{matrix} r_\alpha^* \\ r_\alpha^{**} \end{matrix}\right\} = \frac{1 + \gamma\xi \mp \sqrt{(1+\gamma\xi)^2 - 4(1+\alpha)\gamma\xi}}{2(1+\alpha)\gamma}.$$

(ii) *The sequence $\{t_{\alpha,n}\}$ generated by Newton's method for ϕ_α with initial point $t_{\alpha,0} = 0$ has the closed form*

$$t_{\alpha,n} = \frac{1 - q_\alpha^{2^n-1}}{1 - q_\alpha^{2^n-1}p_\alpha} r_\alpha^* \quad \textit{for each } n = 0, 1, \ldots, \tag{5.19}$$

where

$$q_\alpha := \frac{1-\gamma\xi-\sqrt{(1+\gamma\xi)^2-4(1+\alpha)\gamma\xi}}{1-\gamma\xi+\sqrt{(1+\gamma\xi)^2-4(1+\alpha)\gamma\xi}} \quad \textit{and} \tag{5.20}$$

$$p_\alpha := \frac{1+\gamma\xi-\sqrt{(1+\gamma\xi)^2-4(1+\alpha)\gamma\xi}}{1+\gamma\xi+\sqrt{(1+\gamma\xi)^2-4(1+\alpha)\gamma\xi}}.$$

(iii)

$$\frac{t_{\alpha,n+1}-t_{\alpha,n}}{t_{\alpha,n}-t_{\alpha,n-1}} = \frac{1-q_\alpha^{2^n}}{1-q_\alpha^{2^{n-1}}}\cdot\frac{1-q_\alpha^{2^{n-1}-1}p_\alpha}{1-q_\alpha^{2^{n+1}-1}p_\alpha}q_\alpha^{2^{n-1}} \le {q_\alpha}^{2^n-1}. \tag{5.21}$$

For the following lemma, we define, for $\xi > 0$,

$$I(\xi) = \{\alpha > 0 :\ \xi \le b_\alpha\} = \{\alpha > 0 :\ \gamma\xi \le 1 + 2\alpha - 2\sqrt{\alpha(1+\alpha)}\}. \tag{5.22}$$

Sometimes in order to emphasize the dependence, we write $q(\alpha,\xi)$ for q_α defined by (5.20).

LEMMA 5.5. *The following assertions hold:*

(i) *For each $\alpha > 0$, the function $q(\alpha,\cdot)$ is strictly increasing on $(0, b_\alpha]$.*

(ii) *For each $\xi > 0$, the function $q(\cdot,\xi)$ is strictly increasing on $I(\xi)$.*

Proof. We prove only the assertion (i), as (ii) can be proved similarly. Let $\alpha > 0$. Define

$$g_1(t) = (1+t)^2 - 4(1+\alpha)t \quad \text{for each } t$$

and

$$g_2(t) = 1 - t + \sqrt{g_1(t)} \quad \text{for each } t \in (0, 1+2\alpha-2\sqrt{\alpha(1+\alpha)}].$$

Then,

$$g_1'(t) = 2(1+t) - 4(1+\alpha) \quad \text{for each } t$$

and

$$g_2'(t) = -1 + \frac{g_1'(t)}{2\sqrt{g_1(t)}} = \frac{g_1'(t) - 2\sqrt{g_1(t)}}{2\sqrt{g_1(t)}} \quad \text{for each } t \in (0, 1+2\alpha-2\sqrt{\alpha(1+\alpha)}].$$

Define

$$g(t) = 1 - \frac{2\sqrt{g_1(t)}}{g_2(t)} \quad \text{for each } t \in (0, 1+2\alpha-2\sqrt{\alpha(1+\alpha)}].$$

Then, for each $t \in (0, 1+2\alpha-2\sqrt{\alpha(1+\alpha)}]$,

$$g'(t) = -\frac{g_1'(t)g_2(t) + 2g_1(t) - \sqrt{g_1(t)}g_1'(t)}{g_2^2(t)\sqrt{g_1(t)}} = -\frac{(1-t)g_1'(t) + 2g_1(t)}{g_2^2(t)\sqrt{g_1(t)}}.$$

Since (as can be verified easily)

$$(1-t)g_1'(t) + 2g_1(t) = -4\alpha(1+t) < 0,$$

it follows that $g' > 0$ on $(0, 1 + 2\alpha - 2\sqrt{\alpha(1+\alpha)}]$, and hence g is increasing on $(0, 1 + 2\alpha - 2\sqrt{\alpha(1+\alpha)}]$. Noting that

$$q(\alpha, \xi) = g(\gamma\xi) \quad \text{for each } \xi \in (0, b_\alpha],$$

the desired conclusion holds. The proof is complete. □

THEOREM 5.6. *Let $x_0 \in \mathbb{R}^v$ be a regular point of the inclusion (3.1) with $r > 0$ and $\beta > 0$ such that (4.23) holds. Let $\eta \in [1, +\infty)$, $\Delta \in (0, +\infty]$, $\xi = \eta\beta d(F(x_0), C)$, and $\alpha = \frac{\eta\beta(1-\gamma\xi)^2}{\eta\beta + (1-\eta\beta)(1-\gamma\xi)^2}$. Set*

$$r^*_\alpha = \frac{1 + \gamma\xi - \sqrt{(1+\gamma\xi)^2 - 4(1+\alpha)\gamma\xi}}{2(1+\alpha)\gamma} \quad \text{and} \tag{5.23}$$

$$q_\alpha = \frac{1 - \gamma\xi - \sqrt{(1+\gamma\xi)^2 - 4(1+\alpha)\gamma\xi}}{1 - \gamma\xi + \sqrt{(1+\gamma\xi)^2 - 4(1+\alpha)\gamma\xi}}.$$

Assume that F' satisfies the L-average Lipschitz condition on $\mathbf{B}(x_0, r^*_\alpha)$ and that

$$\xi \le \min\left\{ \frac{1 + 2\eta\beta - 2\sqrt{\eta\beta(1+\eta\beta)}}{\gamma},\ \frac{\left(1 + 2\eta\beta - 2\sqrt{\eta\beta(1+\eta\beta)}\right)(1+\eta\beta)}{1 + \eta\beta - \sqrt{\eta\beta(1+\eta\beta)}} r,\ \Delta \right\}. \tag{5.24}$$

Let $\{x_n\}$ denote the sequence generated by Algorithm A (η, Δ, x_0). *Then $\{x_n\}$ converges at a quadratic rate to some x^* with $F(x^*) \in C$, and the following assertions hold:*

$$\|x_n - x^*\| \le {q_\alpha}^{2^n - 1} r^*_\alpha \quad \text{for all } n = 0, 1, \ldots \tag{5.25}$$

and

$$\|x_{n+1} - x_n\| \le q_\alpha^{2^{n-1}} \|x_n - x_{n-1}\| \quad \text{for all } n = 1, 2, \ldots . \tag{5.26}$$

Proof. By (5.15), $\int_0^\xi L(u)\,\mathrm{d}u = (1-\gamma\xi)^{-2} - 1$; hence α given in the statement of the theorem is consistent with (4.24). Set $\alpha' = \eta\beta$. Then, by (5.16),

$$b_{\alpha'} = \frac{1}{\gamma}\left(1 + 2\eta\beta - 2\sqrt{\eta\beta(1+\eta\beta)}\right) \tag{5.27}$$

and

$$\frac{b_{\alpha'}}{r_{\alpha'}} = \frac{\left(1 + 2\eta\beta - 2\sqrt{\eta\beta(1+\eta\beta)}\right)(1+\eta\beta)}{1 + \eta\beta - \sqrt{\eta\beta(1+\eta\beta)}}.$$

Thus (5.24) reads

$$\xi \le \min\left\{ b_{\alpha'},\ \frac{b_{\alpha'}}{r_{\alpha'}} r,\ \Delta \right\}. \tag{5.28}$$

Since $\gamma\xi < 1$ by (5.24), it is clear from the definition of α that $\alpha < \alpha'$ if $\xi > 0$ and $\alpha = \alpha'$ if $\xi = 0$. Since the function $u \mapsto b_u$ strictly decreasing by Lemma 2.2, it follows that if $\xi > 0$,

$$b_{\alpha'} < b_\alpha. \tag{5.29}$$

We claim that

$$\xi < b_\alpha, \quad \xi \le \Delta, \quad \text{and} \quad r_\alpha^* < r. \tag{5.30}$$

In fact, this claim is trivially true if $\xi = 0$, and so we can assume that $\xi > 0$. Then the first two inequalities follow from (5.28) and (5.29), while the last inequality follows from the fact that $r_\alpha^* < r_{\alpha'}^* \le \frac{r_{\alpha'}}{b_{\alpha'}}\xi \le r$ thanks to (5.28), Lemma 2.1(i), and Lemma 2.2(iii). Therefore, (5.30) is true. Moreover, by (5.30) and (5.16), $\gamma\xi < 1 + 2\alpha - 2\sqrt{\alpha(1+\alpha)}$, the smaller root of the function $t :\to (1+t)^2 - 4(1+\alpha)t$. Therefore, $(1+\gamma\xi)^2 - 4(1+\alpha)\gamma\xi > 0$, and so $q_\alpha < 1$ by (5.23). Now, by Corollary 4.2, the sequence $\{x_n\}$ converges to some x^* with $F(x^*) \in C$, and the following estimates hold for each n:

$$\|x_n - x^*\| \le r_\alpha^* - t_{\alpha,n}, \tag{5.31}$$

$$\begin{aligned} \|x_{n+1} - x_n\| &\le (t_{\alpha,n+1} - t_{\alpha,n}) \left(\frac{\|x_n - x_{n-1}\|}{t_{\alpha,n} - t_{\alpha,n-1}}\right)^2 \\ &\le \left(\frac{t_{\alpha,n+1} - t_{\alpha,n}}{t_{\alpha,n} - t_{\alpha,n-1}}\right) \|x_n - x_{n-1}\|. \end{aligned} \tag{5.32}$$

Hence (5.25) and (5.26) are true because, by (5.19) and (5.21), one has

$$r_\alpha^* - t_{\alpha,n} = \frac{q_\alpha^{2^n-1}(1-p_\alpha)}{1 - q_\alpha^{2^n-1}p_\alpha} r_\alpha^* \le {q_\alpha}^{2^n-1} r_\alpha^* \tag{5.33}$$

and

$$\frac{t_{\alpha,n+1} - t_{\alpha,n}}{t_{\alpha,n} - t_{\alpha,n-1}} \le {q_\alpha}^{2^n-1}. \tag{5.34}$$

Thus the convergence of $\{x_n\}$ is quadratic, and the proof is complete. □

The following result can be proved similarly, but applying Corollary 4.3 in place of Corollary 4.2, and using $\alpha' := \eta\|T_{x_0}^{-1}\|$ (thus $b_{\alpha'} = \frac{1+2\eta\|T_{x_0}^{-1}\| - 2\sqrt{\eta\|T_{x_0}^{-1}\|(1+\eta\|T_{x_0}^{-1}\|)}}{\gamma}$).

THEOREM 5.7. *Let $\eta \in [1, +\infty)$, $\Delta \in (0, +\infty]$, and let C be a cone. Let $x_0 \in \mathbb{R}^v$ such that T_{x_0} carries $\mathbb{R}^v$ onto $\mathbb{R}^m$. Let $\xi = \eta\|T_{x_0}^{-1}\| d(F(x_0), C)$ and*

$$\alpha = \frac{\eta\|T_{x_0}^{-1}\|(1-\gamma\xi)^2}{(\eta-1)\|T_{x_0}^{-1}\| + (1 - (\eta-1)\|T_{x_0}^{-1}\|)(1-\gamma\xi)^2}.$$

Set, as in (5.23),

$$r_\alpha^* = \frac{1 + \gamma\xi - \sqrt{(1+\gamma\xi)^2 - 4(1+\alpha)\gamma\xi}}{2(1+\alpha)\gamma} \quad \textit{and} \tag{5.35}$$

$$q_\alpha = \frac{1 - \gamma\xi - \sqrt{(1+\gamma\xi)^2 - 4(1+\alpha)\gamma\xi}}{1 - \gamma\xi + \sqrt{(1+\gamma\xi)^2 - 4(1+\alpha)\gamma\xi}}.$$

Suppose that F' satisfies the L-average Lipschitz condition on $\mathbf{B}(x_0, r_\alpha^*)$ and that

$$\xi \le \min\left\{\frac{1 + 2\eta\|T_{x_0}^{-1}\| - 2\sqrt{\eta\|T_{x_0}^{-1}\|(1+\eta\|T_{x_0}^{-1}\|)}}{\gamma},\ \Delta\right\}. \tag{5.36}$$

Then the conclusions hold as in Theorem 5.6.

5.3. Extension of Smale's approximate zeros. The following notion of approximate zeros was introduced in [26] for Newton's method. Let f be an operator from a domain D in a Banach space X to another one Y. Recall that Newton's iteration for f is defined as follows:

$$x_{n+1} = x_n - f'(x_n)^{-1} f(x_n), \quad n = 0, 1, \dots . \tag{5.37}$$

The sequence $\{x_n\}$ is said to satisfy Smale's condition if

$$\|x_{n+1} - x_n\| \le \left(\frac{1}{2}\right)^{2^{n-1}} \|x_n - x_{n-1}\| \quad \text{for each } n = 1, 2, \dots . \tag{5.38}$$

Note that (5.38) implies that $\{x_n\}$ is a Cauchy sequence and hence converges (with limit denoted by x^*). By (5.37) it follows that x^* is a zero of f.

DEFINITION 5.8. *Suppose that $x_0 \in D$ is such that Newton iteration* (5.37) *is well defined for f and $\{x_n\}$ satisfies Smale's condition. Then x_0 is said to be an approximate zero of f.*

Note that if x_0 is an approximate zero of f, then Newton iteration (5.37) converges to a zero x^* of f. We now extend the notion of approximate zeros to the Gauss–Newton method for convex composite optimization problems.

DEFINITION 5.9. *Suppose that $x_0 \in D$ is such that the sequence $\{x_n\}$ generated by Algorithm* A (η, Δ, x_0) *converges to a limit x^* solving* (1.1) *and satisfies Smale's condition. Then x_0 is said to be an (η, Δ)-approximate solution of* (1.1).

Recall that L is defined by (5.15).

THEOREM 5.10. *Let $x_0 \in \mathbb{R}^v$ be a regular point of the inclusion* (3.1) *with $r > 0$ and $\beta > 0$ such that* (4.23) *holds. Let $\eta \in [1, +\infty)$, $\Delta \in (0, +\infty]$, $\xi = \eta\beta d(F(x_0), C)$, and*

$$\hat{R} = \left(1 - \sqrt{\frac{\eta\beta}{1+\eta\beta}}\right)\frac{1}{\gamma}.$$

Suppose that F' satisfies the L-average Lipschitz condition on $\mathbf{B}(x_0, \hat{R})$ and that

$$\xi \le \min\left\{\frac{4 + 9\eta\beta - 3\sqrt{\eta\beta(9\eta\beta+8)}}{4\gamma},\ \frac{\left(1 + 2\eta\beta - 2\sqrt{\eta\beta(1+\eta\beta)}\right)(1+\eta\beta)}{1 + \eta\beta - \sqrt{\eta\beta(1+\eta\beta)}} r,\ \Delta\right\}. \tag{5.39}$$

Then, x_0 is an (η, Δ)-approximate solution of (1.1).

Proof. Let α be defined as in Theorem 5.6, and set $\alpha' = \eta\beta$. Then, as in the proof of Theorem 5.6, we have $\alpha \le \alpha'$ and $r_{\alpha'} = \hat{R}$ by (5.16). By Lemma 2.2(iii) and (2.7), it follows that

$$r^*_\alpha \le r^*_{\alpha'} \le r_{\alpha'} = \hat{R}.$$

Thus, by the assumptions, F' satisfies the L-average Lipschitz condition on $\mathbf{B}(x_0, r^*_\alpha)$. On the other hand, noting that

$$\frac{4 + 9\eta\beta - 3\sqrt{\eta\beta(9\eta\beta+8)}}{4\gamma} < \frac{1 + 2\eta\beta - 2\sqrt{\eta\beta(1+\eta\beta)}}{\gamma}, \tag{5.40}$$

we see that (5.39) implies (5.24). Therefore, one can apply Theorem 5.6 to conclude that the sequence $\{x_n\}$ converges to a solution x^* of (1.1) and

$$\|x_{n+1} - x_n\| \le q_\alpha^{2^{n-1}} \|x_n - x_{n-1}\| \quad \text{for all } n = 1, 2, \dots . \tag{5.41}$$

It remains to show that $q_\alpha \le \frac{1}{2}$. To do this we need to emphasize the dependence on the parameters, and so we write $q(\alpha, \xi)$ for q_α defined by (5.20) as before. Note that, by (5.27) the right-hand-side member of the inequality (5.40) is simply $b_{\alpha'}$, while the left-hand-side member majorizes ξ by (5.39). It follows from the monotonicity of $q(\cdot, \cdot)$ established in Lemma 5.5 that

$$q(\alpha, \xi) \le q(\alpha', \xi) \le q\left(\alpha', \frac{4 + 9\eta\beta - 3\sqrt{\eta\beta(9\eta\beta + 8)}}{4\gamma}\right) = \frac{1}{2},$$

where the last equality can be verified elementarily. This completes the proof. ∎

Similar to the above proof, we can use Theorem 5.7 in place of Theorem 5.6 to verify the following result.

THEOREM 5.11. *Let $\eta \in [1, +\infty)$, $\Delta \in (0, +\infty]$, and let C be a cone. Let $x_0 \in \mathbb{R}^v$ such that T_{x_0} carries $\mathbb{R}^v$ onto $\mathbb{R}^m$. Let $\xi = \eta \|T_{x_0}^{-1}\| d(F(x_0), C)$ and*

$$\tilde{R} = \left(1 - \sqrt{\frac{\eta\|T_{x_0}^{-1}\|}{1 + \eta\|T_{x_0}^{-1}\|}}\right) \frac{1}{\gamma}.$$

Suppose that F' satisfies the L-average Lipschitz condition on $\mathbf{B}(x_0, \tilde{R})$ and that

$$\xi \le \min\left\{\frac{4 + 9\eta\|T_{x_0}^{-1}\| - 3\sqrt{\eta\|T_{x_0}^{-1}\|(9\eta\|T_{x_0}^{-1}\| + 8)}}{4\gamma},\ \Delta\right\}. \tag{5.42}$$

Then, x_0 is an (η, Δ)-approximate solution of (1.1).

6. Examples. Let us begin with a simple example demonstrating a quasi-regular point which is not a regular point.

Example 6.1. Consider the operator F from $\mathbb{R}^2$ to $\mathbb{R}^2$ defined by

$$F(x) = \begin{pmatrix} 1 - t_1 + t_2 + t_1^2 \\ 1 - t_1 + t_2 \end{pmatrix} \quad \text{for each } x = (t_1, t_2) \in \mathbb{R}^2,$$

where $\mathbb{R}^2$ is endowed with the l_1-norm. Let $x_0 = 0 \in \mathbb{R}^2$ and $C = \{0\} \subseteq \mathbb{R}^2$. Then

$$F'(x) = \begin{pmatrix} -1 + 2t_1 & 1 \\ -1 & 1 \end{pmatrix} \quad \text{for each } x = (t_1, t_2) \in \mathbb{R}^2;$$

in particular, $F(x_0) = (1, 1)$ and

$$F'(x_0) = \begin{pmatrix} -1 & 1 \\ -1 & 1 \end{pmatrix}.$$

Thus x_0 does not satisfy (3.12). Moreover,

$$\ker F'(x_0) \cap (C - F(x_0))^\ominus = \{(t, t) :\ t \ge 0\} \ne \{0\},$$

and hence x_0 is not a regular point of the inclusion (3.1). In view of the definition of $\mathcal{D}(x)$ in (3.2), we have that, for $x = (t_1, t_2) \in \mathbb{R}^2$,

$$\mathcal{D}(x) = \begin{cases} \{(-\frac{t_1}{2}, -1 + \frac{t_1}{2} - t_2)\}, & t_1 \neq 0, \\ \{(d_1, d_1 - 1 - t_2) : \ d_1 \in \mathbb{R}\}, & t_1 = 0 \end{cases}$$

(note that $F'(x)$ is of full rank if and only if $t_1 \neq 0$). Therefore,

$$d(0, \mathcal{D}(x)) \leq 1 + |t_1| + |t_2| = 1 + \|x\| \quad \text{for each } x = (t_1, t_2) \in \mathbb{R}^2$$

and

$$d(F(x), C) = |1 - t_1 + t_2 + t_1^2| + |1 - t_1 + t_2| \geq 1 - \|x\| \quad \text{for each } x = (t_1, t_2) \in \mathbf{B}(x_0, 1).$$

This implies that

$$d(0, \mathcal{D}(x)) \leq \beta(\|x - x_0\|) d(F(x), C) \quad \text{for each } x = (t_1, t_2) \in \mathbf{B}(x_0, 1),$$

where $\beta(t) = \frac{1+t}{1-t}$ for each $t \in [0, 1)$. Thus, x_0 is a quasi-regular point with quasi-regular radius $\mathbf{r}_{x_0} \geq 1$. In fact, $\mathbf{r}_{x_0} = 1$ because

$$\lim_{t_1 \to 0+} \frac{d(0, \mathcal{D}(x))}{d(F(x), C)} = \lim_{t_1 \to 0+} \frac{1}{t_1} = +\infty$$

as x goes to $(0, -1)$ on the radial $l : 1 - t_1 + t_2 = 0$, $t_1 \geq 0$.

Next we give a few examples to illustrate some situations where our results are applicable but not the earlier results in the literature. For the following examples, recall that C is defined by (1.3) and we take

$$\eta = 1 \quad \text{and} \quad \Delta = +\infty. \tag{6.1}$$

Regarding the convergence issue of Gauss–Newton methods, the advantage of considering Wang's L-average Lipschitz condition rather than the classical Lipschitz condition is shown in the following example, for which Theorem 5.7 is applicable but not Theorem 5.3.

Example 6.2. Let $m = n = 1$ and h be defined by

$$h(y) = \begin{cases} 0, & y \leq 0, \\ y, & y \geq 0. \end{cases}$$

Let τ be a constant satisfying

$$10\sqrt{2} - 14 < \tau < 3 - 2\sqrt{2}, \tag{6.2}$$

and define

$$F(x) = \begin{cases} \tau - x + \frac{x^2}{1-x}, & x \leq \frac{1}{2}, \\ \tau - \frac{1}{2} + 2x^2, & x \geq \frac{1}{2}. \end{cases} \tag{6.3}$$

Then $C = (-\infty, 0]$,

$$F'(x) = \begin{cases} -2 + \frac{1}{(1-x)^2}, & x \leq \frac{1}{2}, \\ 4x, & x \geq \frac{1}{2}, \end{cases}$$

and

$$F''(x) = \begin{cases} \frac{2}{(1-x)^3}, & x < \frac{1}{2}, \\ 4, & x > \frac{1}{2}. \end{cases} \tag{6.4}$$

Let $\gamma = 1$, and let L be defined as in (5.15), that is

$$L(u) = \frac{2}{(1-u)^3} \quad \text{for each } u \text{ with } 0 \le u < 1. \tag{6.5}$$

Then

$$L(u) < L(v) \quad \text{whenever } 0 \le u < v < 1. \tag{6.6}$$

It follows from (6.4) that

$$\sup\{F''(x) : x \in [-r, r] \setminus \{1/2\}\} = \begin{cases} 16, & r \ge \frac{1}{2}, \\ \frac{2}{(1-r)^3}, & 0 < r \le \frac{1}{2}, \end{cases} \tag{6.7}$$

and that

$$0 < F''(u) \le F''(|u|) \le L(|u|) \quad \text{whenever } 1/2 \ne u < 1. \tag{6.8}$$

Let $x_0 = 0$. Then, for all x, $x' \in \mathbf{B}(x_0, 1)$ with $|x'| + |x - x'| < 1$, it follows from (6.6) and (6.8) that

(6.9)

$$|F'(x) - F'(x')| = |x - x'| \int_0^1 F''(x' + t(x - x'))\mathrm{d}t \le |x - x'| \int_0^1 L(|x'| + t|x - x'|)\mathrm{d}t.$$

Thus F' satisfies the L-average Lipschitz condition on $\mathbf{B}(x_0, 1)$ with L defined by (6.5). Note that T_{x_0} carries $\mathbb{R}$ onto $\mathbb{R}$ and $\|T_{x_0}^{-1}\| = 1$ as $F'(x_0) = -1$. Let ξ be defined as in Theorems 5.3 and 5.7. Since $F(x_0) = \tau$ and by (6.2), we have

$$\xi = \|T_{x_0}^{-1}\| d(F(x_0), C) = \tau < 3 - 2\sqrt{2}. \tag{6.10}$$

Thus (5.36) is satisfied. Recalling the definitions of α and r_α^* in Theorem 5.7, we have that $\alpha = 1$ and

$$r_\alpha^* = \frac{1 + \xi - \sqrt{(1+\xi)^2 - 8\xi}}{4} \le \frac{1+\xi}{4} \le \frac{1 + 3 - 2\sqrt{2}}{4} < 1.$$

Therefore Theorem 5.7 is applicable with initial point x_0. We show next that Theorem 5.3 is not applicable here. In fact, by (6.7), one has that, for any $r > 0$, F' is also Lipschitz continuous on $\mathbf{B}(x_0, r)$ with the (least) Lipschitz constant L_r given by

$$L_r = \begin{cases} \frac{2}{(1-r)^3}, & r \le \frac{1}{2}, \\ 16, & r \ge \frac{1}{2}. \end{cases} \tag{6.11}$$

Suppose that there are ξ, L, and R^* satisfying the assumptions stated in Theorem 5.3. For simplicity of notation we write r for R^*. Then by the least property of L_r, (5.12), (5.14), and by a similar argument as for (6.10), we have

$$L \ge L_r, \tag{6.12}$$

$$r = \frac{1 - \sqrt{1 - 2L\xi}}{L}, \tag{6.13}$$

and

$$\tau = \xi \le \frac{1}{2L} \le \frac{1}{2L_r}. \tag{6.14}$$

Since $\tau > 10\sqrt{2} - 14 > \frac{1}{32}$, we have from (6.14) that $L_r < 16$, and it follows from (6.11) that $r < 1/2$ and hence $L_r = \frac{2}{(1-r)^3} \ge 2$. Consequently, by (6.12) and (6.13), we have

$$\tau = \xi = r - \frac{Lr^2}{2} \le r - \frac{L_r r^2}{2} \le r - r^2. \tag{6.15}$$

Combining this with (6.14) and (6.11), we have that

$$\tau \le \min\left\{\frac{(1-r)^3}{4}, r - r^2\right\}. \tag{6.16}$$

Note that the function $r \mapsto \frac{(1-r)^3}{4}$ is decreasing and $r \mapsto r - r^2$ increasing on $[0, \frac{1}{2}]$. Hence

$$\tau \le \min\left\{\frac{(1-r)^3}{4}, r - r^2\right\} = r_0 - r_0^2 = 10\sqrt{2} - 14, \tag{6.17}$$

where $r_0 = 3 - 2\sqrt{2}$ is the least positive root of equation $\frac{(1-r)^3}{4} = r - r^2$. However, (6.17) contradicts (6.2), and therefore Theorem 5.3 is not applicable to x_0.

Even when initial point x_0 is regular, the advantage to considering the quasi regularity bound functions rather than a constant β with the property stated in Proposition 3.3 is shown in the following example, for which Theorem 5.1 (and hence results in [4, 13]) are not applicable while Theorem 5.3 is applicable, which is based on the quasi-regular bound function β_{x_0} satisfying (3.14) rather than the quasi-regular bound constant β given by Proposition 3.3.

Example 6.3. Let $m = n = 1$ and h be defined by

$$h(y) = |y| \quad \text{for each } y \in \mathbb{R}.$$

Then $C = \{0\}$. Let $\frac{\sqrt{3}-1}{4} < \tau \le \frac{1}{4}$ and define

$$F(x) = \tau - x + x^2 \quad \text{for each } x \in \mathbb{R}.$$

Then

$$F'(x) = -1 + 2x \quad \text{for each } x \in \mathbb{R};$$

hence F' is Lipschitz continuous with the modular $L = 2$. Let $x_0 = 0$. It is clear that T_{x_0} carries $\mathbb{R}$ onto $\mathbb{R}$ and $\|T_{x_0}^{-1}\| = 1$ as $F'(x_0) = -1$. Since

$$\|T_{x_0}^{-1}\| d(F(x_0), C) = \tau \le \frac{1}{4}, \tag{6.18}$$

Theorem 5.3 is applicable with initial point $x_0 = 0$. Below we shall show that Theorem 5.1 is not applicable. Suppose, on the contrary, that there exist $r > 0$ and $\beta > 0$ satisfying the assumptions stated in Theorem 5.1 for x_0. Then

$$\mathcal{D}(x) \text{ is nonempty} \quad \text{and} \quad d(0, \mathcal{D}(x)) \le \beta d(F(x), C) \quad \text{for each } x \in \mathbf{B}(x_0, r), \tag{6.19}$$

$$r \geq \frac{1+2\beta\xi-\sqrt{1-(2\beta\xi)^2}}{2\beta}, \tag{6.20}$$

and

$$\beta\tau = \xi \leq \frac{1}{2\beta}. \tag{6.21}$$

By definition, it is easy to see that, for each $x \in \mathbb{R}$,

$$\mathcal{D}(x) = \begin{cases} \{-F'(x)^{-1}F(x)\}, & x \neq \frac{1}{2}, \\ \emptyset, & x = \frac{1}{2}, \end{cases} \tag{6.22}$$

and it follows from (6.19) that $r \leq \frac{1}{2}$ and, for each $x \in \mathbf{B}(x_0, r)$,

$$d(0, \mathcal{D}(x)) = |F'(x)^{-1}F(x)| = \frac{1}{1-2|x|}|F(x)| = \frac{1}{1-2|x|}d(F(x), C). \tag{6.23}$$

By (6.19) this implies that

$$\frac{1}{1-2|x|} \leq \beta \quad \text{for each } x \in \mathbf{B}(x_0, r). \tag{6.24}$$

Considering $x_0 = 0$, this implies $\frac{1}{1-2r} \leq \beta$; that is,

$$2\beta r \leq \beta - 1. \tag{6.25}$$

It follows from (6.20) that

$$1+2\beta\xi-\sqrt{1-(2\beta\xi)^2} \leq \beta - 1, \tag{6.26}$$

or equivalently,

$$((2\xi-1)^2+(2\xi)^2)\beta^2 + 4(2\xi-1)\beta + 3 \leq 0. \tag{6.27}$$

Hence

$$(4(2\xi-1))^2 - 4\cdot 3((2\xi-1)^2+(2\xi)^2) \geq 0, \tag{6.28}$$

which implies that $\xi \leq \frac{\sqrt{3}-1}{4}$. This contradicts the assumption that $\tau > \frac{\sqrt{3}-1}{4}$ because $\xi = \beta\tau \geq \tau$ by (6.21) and (6.24).

We remark that, on one hand, the results in section 5 cover (and improve) the cases considered by Burke and Ferris and by Robinson (the initial points are regular in [4, 18]), and, on the other hand, there are examples of quasi-regular but not regular points x_0 for which Theorem 4.1 is applicable.

Example 6.4. Let $m = n = 3$. To ease our computation, let $\mathbb{R}^3$ be endowed with the l_1-norm. Let h be defined by

$$h(x) = \chi(t_1) + \chi(t_2) + \left|t_3 - t_1 - t_2 - \frac{1}{8}\right| \quad \text{for each } x = (t_1, t_2, t_3) \in \mathbb{R}^3,$$

where $\chi(t)$ is a real-valued function on $\mathbb{R}$ defined by

$$\chi(t) = \begin{cases} -1-t, & t \leq -1, \\ 0, & -1 \leq t \leq 0, \\ t, & t \geq 0. \end{cases}$$

Define

$$A = \{(c_1, c_2, c_3) : \ c_3 = c_1 + c_2\}, \tag{6.29}$$

and let $F : \mathbb{R}^3 \mapsto \mathbb{R}^3$ be defined by

$$F(x) = \begin{pmatrix} \frac{1}{16} - t_1 + t_1^2 + t_2 + t_3 \\ \frac{1}{16} + t_1 - t_2 + t_2^2 + t_3 \\ t_1^2 + t_2^2 + 2t_3 \end{pmatrix} \quad \text{for each } x = (t_1, t_2, t_3) \in \mathbb{R}^3. \tag{6.30}$$

Then

$$C = \left\{(c_1, c_2, c_3) : \ c_1, c_2 \in [-1, 0], \, c_3 = c_1 + c_2 + \frac{1}{8}\right\}, \tag{6.31}$$

$$C - F(x) \text{ is contained in } A, \tag{6.32}$$

$$F'(x) = \begin{pmatrix} -1 + 2t_1 & 1 & 1 \\ 1 & -1 + 2t_2 & 1 \\ 2t_1 & 2t_2 & 2 \end{pmatrix} \quad \text{for each } x = (t_1, t_2, t_3) \in \mathbb{R}^3, \tag{6.33}$$

and hence

$$C - F(x) \text{ is contained in } A = \{F'(x)d : \ d \in \mathbb{R}^3\} \quad \text{for each } x \in \mathbb{R}^3. \tag{6.34}$$

In particular, for $x_0 = 0$, we have that

$$F'(x_0) = \begin{pmatrix} -1 & 1 & 1 \\ 1 & -1 & 1 \\ 0 & 0 & 2 \end{pmatrix}, \tag{6.35}$$

and hence

$$\ker F'(x_0) = \{(t, t, 0) : \ t \in \mathbb{R}\}. \tag{6.36}$$

Since $F(x_0) = (\frac{1}{16}, \frac{1}{16}, 0)$, one has that

$$\ker F'(x_0) \cap (C - F(x_0))^{\ominus} = \{(t, t, 0) : \ t \geq 0\}; \tag{6.37}$$

hence x_0 is not a regular point of (3.1), and the condition of Robinson is not satisfied (see Proposition 3.7). Below we shall show that x_0 is a quasi-regular point with the quasi-regular radius $\mathbf{r}_{x_0}$ and the quasi-regular bound function β_{x_0} satisfying respectively

$$\mathbf{r}_{x_0} \geq \frac{3}{4} \quad \text{and} \quad \beta_{x_0}(t) \leq \frac{2}{3 - 4t} \quad \text{for each } t \in \left[0, \frac{3}{4}\right). \tag{6.38}$$

To do this, we note first that F' satisfies the L-average Lipschitz condition on $\mathbb{R}^3$ with $L = 2$,

$$\|F'(x) - F'(y)\| \leq 2\|x - y\| \quad \text{for each } x, \, y \in \mathbb{R}^3, \tag{6.39}$$

and the rank of $F'(x)$ is given by

$$\operatorname{rank} F'(x) = 2 \quad \text{for each } x \in \mathbb{R}^3. \tag{6.40}$$

Since

$$F'(x) = F'(x_0) + F'(x) - F'(x_0) \quad \text{and} \quad \|F'(x) - F'(x_0)\| \le 2\|x - x_0\|,$$

it follows from the perturbation property of matrixes (cf. [27, 30]) that

$$\|F'(x)^\dagger\| \le \frac{\|F'(x_0)^\dagger\|}{1 - 2\|x - x_0\|\|F'(x_0)^\dagger\|} \tag{6.41}$$

holds for each $x \in \mathbb{R}^3$ with $2\|x - x_0\|\|F'(x_0)^\dagger\| < 1$, where $A^\dagger$ denotes the Moore–Penrose generalized inverse of the matrix A (cf. [27, 30]). By (6.35), one has that

$$F'(x_0)^\dagger = \begin{pmatrix} -\frac{1}{4} & \frac{1}{4} & 0 \\ \frac{1}{4} & -\frac{1}{4} & 0 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{3} \end{pmatrix} \quad \text{and} \quad \|F'(x_0)^\dagger\| = \frac{2}{3}. \tag{6.42}$$

This together with (6.41) implies that

$$\|F'(x)^\dagger\| \le \frac{2}{3 - 4\|x - x_0\|} \quad \text{for each } x \in \mathbf{B}\left(x_0, \frac{3}{4}\right). \tag{6.43}$$

On the other hand, by (6.33) and (3.2), we have

$$\mathcal{D}(x) = F'(x)^\dagger(C - F(x)) \quad \text{for each } x \in \mathbb{R}^3 \tag{6.44}$$

and, consequently, for each $x \in \mathbf{B}(x_0, \frac{3}{4})$,

$$d(0, \mathcal{D}(x)) \le \|F'(x)^\dagger\| d(F(x), C) \le \frac{2}{3 - 4\|x - x_0\|} d(F(x), C). \tag{6.45}$$

This shows that x_0 is a quasi-regular point with the quasi-regular radius $\mathbf{r}_{x_0}$ and the quasi-regular bound function β_{x_0} satisfying (6.38). Let $\mathbf{r} = \frac{3}{4}$. Recalling (4.2) and (6.1), it follows from (6.38) that

$$\alpha_0(\mathbf{r}) \le \sup\left\{\frac{\beta(t)}{\beta(t)2t + 1} : \xi \le t < \mathbf{r}\right\} = \frac{2}{3}, \tag{6.46}$$

where $\beta(t) := \frac{2}{3-4t}$ for each $t \in [0, \frac{3}{4})$. Thus taking $\alpha = \frac{2}{3}$ in (2.1), we get that

$$r_\alpha = \frac{3}{4} \quad \text{and} \quad b_\alpha = \frac{3}{8}. \tag{6.47}$$

By (4.1) and (6.38),

$$\xi = \beta_{x_0}(0) d(F(x_0), C) \le \beta(0)\|F(x_0) - c_0\| = \frac{1}{6},$$

where $c_0 = (0, 0, \frac{1}{8})$. It follows that (4.4) is satisfied. Hence Theorem 4.1 is applicable with initial point x_0 even though it is not a regular point.

7. Conclusion. In connection with inclusion problem (1.2) and for a given point x_0 we introduce two new notions: (a) the L-average Lipschitz condition for F' and (b) quasi regularity (with the associate quasi-regular radius r_{x_0} and quasi-regular bound function β_{x_0}). The notion (a) extends the classical Lipschitz condition and Smale's condition, and notion (b) extends the regularity. When Robinson's condition (3.12) is satisfied, x_0 is shown to be a regular point, and the associate quasi-regular radius r_{x_0} as well as the associate quasi-regular bound function β_{x_0} are estimated if in addition F' satisfies (a) with suitable L. We provide sufficient conditions for convergence results with a quasi-regular initial point x_0 in the Gauss–Newton method for the convex composition optimization problem (1.1) with C given to be the set of all minimizers of a convex function h. These conditions are given in terms of r_{x_0}, β_{x_0}, and L in (a). Examples are given to show that the new concept and results are nontrivial extensions of the existing ones.

Acknowledgment. We thank the referees for suggestions which helped our presentation.

REFERENCES

[1] J. M. Borwein, *Stability and regular points of inequality systems*, J. Optim. Theory Appl., 48 (1986), pp. 9–52.
[2] J. V. Burke, *Descent methods for composite nondifferentiable optimization problems*, Math. Programming, 33 (1985), pp. 260–279.
[3] J. V. Burke, *An exact penalization viewpoint of constrained optimization*, SIAM J. Control Optim., 29 (1991), pp. 968–998.
[4] J. V. Burke and M. C. Ferris, *A Gauss–Newton method for convex composite optimization*, Math. Programming, 71 (1995), pp. 179–194.
[5] J. V. Burke and R. A. Poliquin, *Optimality conditions for non-finite valued convex composite function*, Math. Programming, 57 (1992), pp. 103–120.
[6] R. Flecher, *Second order correction for nondifferentiable optimization*, in Numerical Analysis, G. A. Watson, ed., Lecture Notes in Math. 912, Springer, Berlin, 1982, pp. 85–114.
[7] R. Flecher, *Practical Methods of Optimization*, 2nd ed., Wiley, New York, 1987.
[8] W. B. Gragg and R. A. Tapai, *Optimal error bounds for the Newton–Kantorovich theorems*, SIAM J. Numer. Anal., 11 (1974), pp. 10–13.
[9] J. Hiriart-Urruty and C. Lemarechal, *Convex Analysis and Minimization Algorithms* II. *Advanced Theory and Bundle Methods*, Grundlehren Math. Wiss. 305, Springer, New York, 1993.
[10] K. Jittorntrum and M. R. Osborne, *Strong uniqueness and second order convergence in nonlinear discrete approximation*, Numer. Math., 34 (1980), pp. 439–455.
[11] L. V. Kantorovich, *On Newton method for functional equations*, Dokl. Akad. Nauk USSR, 59 (1948), pp. 1237–1240.
[12] L. V. Kantorovich and G. P. Akilov, *Functional Analysis*, Pergamon Press, New York, 1982.
[13] C. Li and X. H. Wang, *On convergence of the Gauss–Newton method for convex composite optimization*, Math. Program., 91 (2002), pp. 349–356.
[14] W. Li and I. Singer, *Global error bounds for convex multifunctions and applications*, Math. Oper. Res., 23 (1998), pp. 443–462.
[15] K. Madsen, *Minimization of Nonlinear Approximation Function*, Ph.D. thesis, Institute of Numerical Analysis, Technical University of Denmark, Lyngby, Denmark, 1985.
[16] K. F. Ng and X. Y. Zheng, *Characterizations of error bounds for convex multifunctions on Banach spaces*, Math. Oper. Res., 29 (2004), pp. 45–63.
[17] A. M. Ostrowski, *Solutions of Equations in Euclidean and Banach Spaces*, Academic Press, New York, 1973.
[18] S. M. Robinson, *Extension of Newton's method to nonlinear functions with values in a cone*, Numer. Math., 19 (1972), pp. 341–347.
[19] S. M. Robinson, *Normed convex process*, Trans. Amer. Math. Soc., 174 (1972), pp. 127–140.
[20] S. M. Robinson, *Stability theory for systems of inequalities, Part* I: *Linear systems*, SIAM J. Numer. Anal., 12 (1975), pp. 754–769.

[21] S. M. Robinson, *Stability theory for systems of inequalities, Part* II: *Differentiable nonlinear systems*, SIAM J. Numer. Anal., 13 (1976), pp. 497–513.
[22] R. T. Rockafellar, *First and second order epi-differentiability in nonlinear programming*, Trans. Amer. Math. Soc., 307 (1988), pp. 75–108.
[23] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[24] R. T. Rockafellar, *Monotone Processes of Convex and Concave Type*, Mem. Amer. Math. Soc. 77, AMS, Providence, RI, 1967.
[25] R. T. Rockafellar, *First- and second-order epi-differentiability in nonlinear programming*, Trans. Amer. Math. Soc., 307 (1988), pp. 75–108.
[26] S. Smale, *Newton's method estimates from data at one point*, in The Merging of Disciplines: New Directions in Pure, Applied and Computational Mathematics, R. Ewing, K. Gross, and C. Martin, eds., Springer, New York, 1986, pp. 185–196.
[27] G. W. Stewart, *On the continuity of the generalized inverse*, SIAM J. Appl. Math., 17 (1969), pp. 33–45.
[28] X. H. Wang, *Convergence of Newton's method and inverse function theorem in Banach space*, Math. Comp., 68 (1999), pp. 169–186.
[29] X. H. Wang, *Convergence of an iteration process*, Kexue Tongbao, 20 (1975), pp. 558–559.
[30] P. Wedin, *Perturbation theory for pseudo-inverse*, BIT, 13 (1973), pp. 217–232.
[31] R. S. Womersley, *Local properties of algorithms for minimizing nonsmooth composite function*, Math. Programming, 32 (1985), pp. 69–89

SIAM J. OPTIM.
Vol. 18, No. 2, pp. 643–665

THE SECQ, LINEAR REGULARITY, AND THE STRONG CHIP FOR AN INFINITE SYSTEM OF CLOSED CONVEX SETS IN NORMED LINEAR SPACES*

CHONG LI†, K. F. NG‡, AND T. K. PONG‡

Abstract. We consider a (finite or infinite) family of closed convex sets with nonempty intersection in a normed space. A property relating their epigraphs with their intersection's epigraph is studied, and its relations to other constraint qualifications (such as the linear regularity, the strong CHIP, and Jameson's (G)-property) are established. With suitable continuity assumption we show how this property can be ensured from the corresponding property of some of its finite subfamilies.

Key words. system of closed convex sets, interior-point condition, strong conical hull intersection property

AMS subject classifications. Primary, 90C34, 90C25; Secondary, 52A05, 41A29

DOI. 10.1137/060652087

1. Introduction.

In dealing with a lower semicontinuous extended real-valued function ϕ defined on a Banach space (or more generally, a normed linear space) X, it is not only natural but also useful to study its relation with the epigraph $\operatorname{epi}\phi := \{(x,r) \in X \times \mathbb{R} : \phi(x) \le r\}$, which is clearly a closed convex subset of the product $X \times \mathbb{R}$. Conversely, given a nonempty closed convex set C in X, let σ_C denote the support function of C, which is defined by

$$\sigma_C(x^*) = \sup\{\langle x^*, x\rangle : x \in C\}, \qquad x^* \in X^*,$$

where X^* denotes the dual space of X and $\langle x^*, x\rangle = x^*(x)$, the value of the functional x^* at x. Thus σ_C is a w^*-lower semicontinuous convex function and $\operatorname{epi}\sigma_C$ is a w^*-closed convex subset of $X^* \times \mathbb{R}$. In this paper, we shall apply this simple duality between C and $\operatorname{epi}\sigma_C$ to study several important aspects (including the regularity, the strong conical hull intersection property (CHIP), Jameson's property (G), and other constraint qualifications) for a *CCS*-system $\{C_i : i \in I\}$, by which we mean a family of closed convex sets in X with nonempty intersection $\cap_{i\in I} C_i$, where I is an index set.

For the case when I is finite, the concept of regularity and its quantitative versions were introduced in [4, 5, 6] by Bauschke, Borwein, and Li and were utilized to establish norm or linear convergence results. The concept of the strong CHIP was introduced by Deutsch, Li, and Ward in [12], and was utilized in [13], as well as in [9, 24, 25], to reformulate certain optimization problems with constraints. All the works cited above were in the Hilbert space or Euclidean space setting. The concept of property (G) was introduced by Jameson [17] for a pair of cones, and was utilized to give a duality

*Received by the editors February 13, 2006; accepted for publication (in revised form) January 9, 2007; published electronically August 8, 2007.

http://www.siam.org/journals/siopt/18-2/65208.html

†Department of Mathematics, Zhejiang University, Hangzhou 310027, People's Republic of China (cli@zju.edu.cn). This author was supported in part by the National Natural Science Foundation of China (grant 10671175) and Program for New Century Excellent Talents in University.

‡Department of Mathematics, Chinese University of Hong Kong, Hong Kong, People's Republic of China (kfng@math.cuhk.edu.hk, tkpong@gmail.com). K. F. Ng was supported by a direct grant (CUHK) and an Earmarked Grant from the Research Grant Council of Hong Kong.

characterization of the linear regularity. In improving the partial results obtained by Lewis and Pang (see [23, 31]) and by Bauschke, Borwein, and Li [6], Jameson's result was extended by Ng and Yang [30] to the general case (without the additional assumption that each C_i is a cone), but still only for finite I. For the case when X is a Hilbert space, the same result was also independently obtained by Bakan, Deutsch, and Li in [3].

In this paper, we extend the above mentioned results to cover the case when I is infinite. From both theoretical and application points of view, the extension from the finite case to the infinite one is of importance. Regarding the strong CHIP, such an extension has already been done rather successfully with many interesting applications (see, for example, [27, 28]). Our investigation is made through the consideration of epigraphs, and in particular by virtue of that, of a new constraint qualification, defined below. Our works in this connection are inspired by the recent works of Jeyakumar and his collaborates (see [7, 18, 19, 20, 22], for example), who made use of epigraphs to provide sufficient conditions to ensure the strong CHIP (for a finite collection of closed convex sets), and study systems of convex inequalities. We say that a *CCS*-system $\{C_i : i \in I\}$ satisfies the SECQ (sum of epigraphs constraint qualification) if

$$\operatorname{epi} \sigma_{\cap_{i\in I} C_i} = \sum_{i\in I} \operatorname{epi} \sigma_{C_i}.$$

In section 4, we study the interrelationship between this property and other constraint qualifications, especially the linear regularity. Also, since this property is stronger than the strong CHIP (and the converse holds in some important cases; see Theorem 3.1), it is both natural and useful to inquire whether or not the sufficient conditions originally provided to ensure the strong CHIP can in fact ensure the SECQ. In this connection, let us recall the following results proved in [27] (see, in particular, Theorems 4.1 and 5.1 therein). For the remainder of this section, we assume that I is a compact metric space (needless to say, if I is finite, then it is compact under the discrete metric) and see the next section for definitions of the undefined terms.

THEOREM 1.1. *Consider the CCS-system $\{D, C_i : i \in I\}$. Suppose that*

(a) *D is of finite dimension;*

(b) *the set-valued map $i \mapsto (\operatorname{aff} D) \cap C_i$ is lower semicontinuous on I;*

(c) *there exist $x_0 \in D \cap (\cap_{i\in I} C_i)$ and $r > 0$ such that*

$$(\operatorname{aff} D) \cap B(x_0, r) \subseteq C_i \quad \text{for each } i \in I; \tag{1.1}$$

(d) *the pair $\{\operatorname{aff} D, C_i\}$ has the strong CHIP for each $i \in I$.*

Then $\{D, C_i : i \in I\}$ has the strong CHIP.

THEOREM 1.2. *Consider the CCS-system $\{D, C_i : i \in I\}$. Suppose that*

(a) *D is of finite dimension l;*

(b) *the set-valued map $i \mapsto (\operatorname{aff} D) \cap C_i$ is lower and upper semicontinuous on I;*

(c) *for any finite subset J of I with number of elements $|J| \leq l$, there exist $x_0 \in D$ and $r > 0$ such that*

$$(\operatorname{aff} D) \cap B(x_0, r) \subseteq C_i \quad \text{for each } i \in J;$$

(d) *for any finite subset J of I, the subsystem $\{D, C_j : j \in J\}$ has the strong CHIP.*

Then $\{D, C_i : i \in I\}$ *has the strong CHIP.*

In section 5, we present corresponding results for the SECQ, and as a consequence Theorems 1.1 and 1.2 are recaptured with some significant improvements. In our Corollary 5.5, condition (c) in Theorem 1.1 can be considerably weakened to require (1.1) to hold for each $i \in J$ with some finite subsets J of I and to allow r to depend on J. In our Corollary 5.6, we show that the words "and upper" in Theorem 1.2(b) can be dropped and that (d) can be weakened to require the strong CHIP to hold only for subsystems $\{D, C_j : j \in J\}$ with $|J| = l + 1$.

2. Notations and preliminary results. The notations used in the present paper are standard (cf. [8, 15]). In particular, we assume throughout the whole paper that X is a real normed linear space (we remark that some results in this section hold for general locally convex spaces). We use $\mathbf{B}(x, \epsilon)$ to denote the closed ball with center x and radius ϵ. For a set A in X (or in $\mathbb{R}^n$), the interior (resp., relative interior, closure, convex hull, convex cone hull, linear hull, affine hull, boundary) of A is denoted by $\operatorname{int} A$ (resp., $\operatorname{ri} A$, $\overline{A}$, $\operatorname{co} A$, $\operatorname{cone} A$, $\operatorname{span} A$, $\operatorname{aff} A$, $\operatorname{bd} A$), and the negative polar cone $A^{\ominus}$ is the set defined by

$$A^{\ominus} = \{x^* \in X^* \ : \ \langle x^*, z\rangle \le 0 \ \text{ for all } z \in A\},$$

which coincides with the polar A° of A when A is a cone. The normal cone of A at z_0 is denoted by $N_A(z_0)$ and defined by $N_A(z_0) = (A - z_0)^{\ominus}$. Let Z be a closed convex nonempty subset of X. The interior and the boundary of A relative to Z are, respectively, denoted by $\operatorname{rint}_Z A$ and $\operatorname{bd}_Z A$; they are defined to be, respectively, the interior and the boundary of the set $\operatorname{aff} Z \cap A$ in the metric space $\operatorname{aff} Z$. Thus, a point $z \in \operatorname{rint}_Z A$ if and only if there exists $\varepsilon > 0$ such that

$$z \in (\operatorname{aff} Z) \cap \mathbf{B}(z, \varepsilon) \subseteq A,$$

while $z \in \operatorname{bd}_Z A$ if and only if $z \in \operatorname{aff} Z$ and, for any $\varepsilon > 0$, $(\operatorname{aff} Z) \cap \mathbf{B}(z, \varepsilon)$ intersects A and its complement.

For a closed subset A of X, the indicator function δ_A and the support function σ_A of set A are, respectively, defined by

$$\delta_A(x) := \begin{cases} 0, & x \in A, \\ \infty, & \text{otherwise} \end{cases}$$

and

$$\sigma_A(x^*) := \sup_{x \in A} \langle x^*, x\rangle \quad \text{for each } x^* \in X^*.$$

Let f be a proper lower semicontinuous extended real-valued function on X. The domain of f is denoted by $\operatorname{dom} f := \{x \in X : f(x) < +\infty\}$. Then the subdifferential of f at $x \in \operatorname{dom} f$, denoted by $\partial f(x)$, is defined by

$$\partial f(x) := \{z^* \in X^* : \ f(x) + \langle z^*, y - x\rangle \le f(y) \quad \text{for all } y \in X\}.$$

Let f, g be proper functions, respectively, defined on X and X^*. Let f^*, g^* denote their conjugate functions, that is,

$$f^*(x^*) := \sup\{\langle x^*, x\rangle - f(x) : \ x \in X\} \quad \text{for each } x^* \in X^*,$$
$$g^*(x) := \sup\{\langle x^*, x\rangle - g(x^*) : \ x^* \in X^*\} \quad \text{for each } x \in X.$$

The epigraph of a function f on X is denoted by $\operatorname{epi} f$ and defined by

$$\operatorname{epi} f := \{(x,r) \in X \times \mathbb{R} :\ f(x) \le r\}.$$

Then, for proper lower semicontinuous extended real-valued convex functions f_1 and f_2 on X, we have

$$f_1 \le f_2 \Longleftrightarrow f_1^* \ge f_2^* \Longleftrightarrow \operatorname{epi} f_1^* \subseteq \operatorname{epi} f_2^*, \tag{2.1}$$

where the forward direction of the first arrow and the second equivalence are easy to verify, while the backward direction of the first arrow is standard (cf. [35, Theorem 2.3.3]).

For closed convex sets A, B, the following assertions are well known and easy to verify:

$$\sigma_A = \delta_A^*, \tag{2.2}$$

$$N_A(x) = \partial \delta_A(x) \quad \text{for each } x \in A, \tag{2.3}$$

$$\sigma_A(x^*) = \langle x^*, x\rangle \Leftrightarrow x^* \in N_A(x) \Longleftrightarrow (x^*,\, \langle x^*, x\rangle) \in \operatorname{epi} \sigma_A \quad \text{for each } x \in A, \tag{2.4}$$

and

$$\operatorname{epi} \sigma_A \subseteq \operatorname{epi} \sigma_B \quad \text{if } A \supseteq B.$$

Let $\{A_i :\ i \in J\}$ be a family of subsets of X containing the origin. The set $\sum_{i\in J} A_i$ is defined by

$$\sum_{i\in J} A_i = \begin{cases} \left\{\sum_{i\in J_0} a_i :\quad a_i \in A_i, \quad \emptyset \ne J_0 \subseteq J \text{ being finite}\right\} & \text{if } J \ne \emptyset, \\ \{0\} & \text{if } J = \emptyset. \end{cases}$$

Let I be an arbitrary index set. The following concept of the strong CHIP plays an important role in optimization theory (see [3, 6, 9, 10, 11, 33]) and is due to [12, 13] in the case when I is finite and to [26, 27] in the case when I is infinite.

DEFINITION 2.1. *Let $\{C_i :\ i \in I\}$ be a collection of convex subsets of X. The collection is said to have*

(a) *the strong CHIP at $x \in \cap_{i\in I} C_i$ if $N_{\cap_{i\in I} C_i}(x) = \sum_{i\in I} N_{C_i}(x)$, that is,*

$$\left(\bigcap_{i\in I} C_i - x\right)^{\ominus} = \sum_{i\in I} (C_i - x)^{\ominus};$$

(b) *the strong CHIP if it has the strong CHIP at each point of $\cap_{i\in I} C_i$;*

(c) *the SECQ if $\operatorname{epi} \sigma_{\cap_{i\in I} C_i} = \sum_{i\in I} \operatorname{epi} \sigma_{C_i}$.*

Note that $N_{\cap_{i\in I} C_i}(x) \supseteq \sum_{i\in I} N_{C_i}(x)$ holds automatically for $x \in \cap_{i\in I} C_i$. Hence $\{C_i :\ i \in I\}$ has the strong CHIP at x if and only if

$$N_{\cap_{i\in I} C_i}(x) \subseteq \sum_{i\in I} N_{C_i}(x).$$

To establish a similar property regarding the SECQ, we first need to extend [16, part X, Theorem 2.4.4] to the setting of normed linear spaces. We recall that for an arbitrary function f defined on X^*, we define $\overline{\operatorname{co}} f^{w^*}$ by (cf. [35, page 63])

$$\operatorname{epi}(\overline{\operatorname{co}} f^{w^*}) := \overline{\operatorname{co}(\operatorname{epi} f)}^{w^*}.$$

LEMMA 2.2. *Let $\{g_i : i \in I\}$ be a family of proper convex lower semicontinuous functions on X with $\sup_{i\in I} g_i(x_0) < +\infty$ for some $x_0 \in X$. Then for all $y^* \in X^*$, $(\sup_{i\in I} g_i)^*(y^*) = \overline{\mathrm{co}(\inf_{i\in I}(g_i^*))}^{w^*}(y^*)$.*

Proof. It is well known (and immediate from the definition of the conjugate) that for a family of proper convex lower semicontinuous functions $(f_i)_{i\in I}$ on X,

$$\left(\inf_{i\in I} f_i\right)^* = \sup_{i\in I} f_i^*. \tag{2.5}$$

Now, since g_i is proper convex lower semicontinuous for each $i \in I$, $g_i^{**} = g_i$ and g_i^* is proper (cf. [35, Theorem 2.3.3]). Applying (2.5) to g_i^* in place of f_i, we see that

$$\left(\inf_{i\in I} g_i^*\right)^* = \sup_{i\in I} g_i^{**} = \sup_{i\in I} g_i.$$

From this and the properness assumption on $\sup_{i\in I} g_i$, we obtain from [35, Theorem 2.3.4] that $(\inf_{i\in I} g_i^*)^{**} = \overline{\mathrm{co}(\inf_{i\in I} g_i^*)}^{w^*}$. The result follows. □

The following lemma was stated without proof in [21, page 902]. We give a proof here for the sake of completeness (note that the condition that "$\sup_{i\in I} g_i$ is proper" is needed).

LEMMA 2.3. *Let $\{g_i : i \in I\}$ be a system of proper convex lower semicontinuous functions on X with $\sup_{i\in I} g_i(x_0) < +\infty$ for some $x_0 \in X$. Then*

$$\mathrm{epi}\left(\sup_{i\in I} g_i\right)^* = \overline{\mathrm{co}\bigcup_{i\in I} \mathrm{epi}\, g_i^*}^{w^*}. \tag{2.6}$$

Proof. For the family $\{g_i^* : i \in I\}$ of proper convex lower semicontinuous functions on X we have $\cup_{i\in I} \mathrm{epi}\, g_i^* \subset \mathrm{epi}\,(\inf_{i\in I} g_i^*) \subset \overline{\mathrm{co}\cup_{i\in I} \mathrm{epi}\, g_i^*}^{w^*}$. This implies that $\overline{\mathrm{epi}\,(\inf_{i\in I} g_i^*)}^{w^*} = \overline{\mathrm{co}\cup_{i\in I} \mathrm{epi}\, g_i^*}^{w^*}$. The conclusion follows on invoking Lemma 2.2. □

PROPOSITION 2.4. *Let $\{C_i : i \in I\}$ be a collection of closed convex sets in X with $C := \cap_{i\in I} C_i \neq \emptyset$. Then*

$$\mathrm{epi}\,\sigma_C = \overline{\sum_{i\in I} \mathrm{epi}\,\sigma_{C_i}}^{w^*}.$$

Proof. Note that $\sup_{i\in I} \delta_{C_i} = \delta_C$ and that $\sigma_C = \delta_C^*$ by (2.2). It follows that $\mathrm{epi}\,\sigma_C = \mathrm{epi}(\sup_{i\in I} \delta_{C_i})^*$. Consequently, by (2.6) and (2.2), one has that

$$\mathrm{epi}\,\sigma_C = \overline{\mathrm{co}\bigcup_{i\in I} \mathrm{epi}\,\delta_{C_i}^*}^{w^*} = \overline{\mathrm{co}\bigcup_{i\in I} \mathrm{epi}\,\sigma_{C_i}}^{w^*} = \overline{\sum_{i\in I} \mathrm{epi}\,\sigma_{C_i}}^{w^*},$$

where the last equality holds because $\mathrm{epi}\,\sigma_{C_i}$ is clearly a cone for each $i \in I$. □

COROLLARY 2.5. *Let $\{C_i : i \in I\}$ be a collection of closed convex sets in X with $C := \bigcap_{i\in I} C_i \neq \emptyset$. Then the following equivalences are true:*

$$\begin{aligned} &\{C_i : i \in I\} \text{ satisfies the SECQ} \iff \textstyle\sum_{i\in I} \mathrm{epi}\,\sigma_{C_i} \text{ is } w^*\text{-closed} \\ &\iff \mathrm{epi}\,\sigma_C \subseteq \sum_{i\in I} \mathrm{epi}\,\sigma_{C_i}. \end{aligned} \tag{2.7}$$

The following simple proposition states that the SECQ is invariant under translation.

PROPOSITION 2.6. *Let $\{C_i : i \in I\}$ be a family of closed convex sets in X. Suppose that $C := \bigcap_{i\in I} C_i \neq \emptyset$. Then $\{C_i : i \in I\}$ satisfies the SECQ if and only if the system $\{C_i - x : i \in I\}$ does for each $x \in X$.*

Proof. Let $x \in X$. Note that

$$(y^*, \alpha) \in \operatorname{epi} \sigma_{C-x} \Longleftrightarrow (y^*, \alpha + \langle y^*, x\rangle) \in \operatorname{epi} \sigma_C$$

and

$$(y^*, \alpha) \in \sum_{i\in I} \operatorname{epi} \sigma_{C_i - x} \Longleftrightarrow (y^*, \alpha + \langle y^*, x\rangle) \in \sum_{i\in I} \operatorname{epi} \sigma_{C_i}.$$

Hence the conclusion follows from Corollary 2.5. □

We will need the following notion of semicontinuity of set-valued maps in sections 4 and 5. Readers may refer to standard texts such as [1, 32].

DEFINITION 2.7. *Let Q be a metric space. Let X be a normed linear space and let $t_0 \in Q$. A set-valued function $F : Q \to 2^Y \setminus \{\emptyset\}$ is said to be*

(i) *lower semicontinuous at t_0 if, for any $y_0 \in F(t_0)$ and any $\epsilon > 0$, there exists a neighborhood $U(t_0)$ of t_0 such that $\mathbf{B}(y_0, \epsilon) \cap F(t) \neq \emptyset$ for each $t \in U(t_0)$;*

(ii) *lower semicontinuous on Q if it is lower semicontinuous at each $t \in Q$.*

The following characterization regarding the lower semicontinuity is a reformulation of the equivalence of (i) and (ii) in [27, Proposition 3.1]. For a closed convex set S in a normed linear space X, let $d_S(\cdot)$ denote the distance function of S defined by $d_S(x) = \inf\{\|x - y\| : y \in S\}$ for each $x \in X$. Furthermore, let $\liminf_{t\to t_0} F(t)$ denote the lower limit of the set-valued function F at $t_0 \in Q$ which is defined by

$$\liminf_{t\to t_0} F(t) := \{z \in X : \exists\{z_t\}_{t\in Q} \text{ with } z_t \in F(t) \text{ such that } z_t \to z \text{ as } t \to t_0\}.$$

PROPOSITION 2.8. *Let Q be a metric space. Let $F : Q \to 2^X \setminus \{\emptyset\}$ be a set-valued function and let $t_0 \in Q$. Then the following statements are equivalent:*

(i) *F is lower semicontinuous at t_0.*

(ii) *For any $y_0 \in F(t_0)$, $\lim_{t\to t_0} d_{F(t)}(y_0) = 0$.*

(iii) *$F(t_0) \subseteq \liminf_{t\to t_0} F(t)$.*

We collect some properties of the lower limit of the set-valued function F at $t_0 \in Q$ in the following proposition. The first property is direct from the definition and the second property is a direct consequence of [32, Proposition 4.15].

PROPOSITION 2.9. *Let Q be a metric space and X a normed linear space. Let $F : Q \to 2^X \setminus \{\emptyset\}$ be a set-valued function such that $F(t)$ is convex for each $t \in Q$. Let $t_0 \in Q$. Then $\liminf_{t\to t_0} F(t)$ is convex.*

Moreover, if X is finite dimensional and B is a compact subset contained in $\operatorname{int}(\liminf_{t\to t_0} F(t))$ (e.g., F is lower semicontinuous and B is a compact set contained in $\operatorname{int}(F(t_0))$), then there exists a neighborhood $U(t_0)$ of t_0 such that $B \subseteq \operatorname{int} F(t)$ for each $t \in U(t_0)$.

3. The strong CHIP and the SECQ. Recall that I is an arbitrary index set and $\{C_i : i \in I\}$ is a collection of nonempty closed convex subsets of X. We denote $\bigcap_{i\in I} C_i$ by C and assume that $0 \in C$ throughout the whole paper. The following theorem describes a relationship between the strong CHIP and the SECQ for the system $\{C_i : i \in I\}$. Results in this section are folklore; we include their proofs here for the sake of completeness.

THEOREM 3.1. *If $\{C_i : i \in I\}$ satisfies the SECQ, then it has the strong CHIP; the converse conclusion holds if* $\operatorname{dom}\sigma_C \subseteq \operatorname{Im}\partial\,\delta_C$, *that is, if*

$$\operatorname{dom}\sigma_C \subseteq \bigcup_{x\in C} N_C(x). \tag{3.1}$$

Proof. Suppose that $\{C_i : i \in I\}$ satisfies the SECQ. Let $x \in C$ and $y^* \in N_C(x)$. Then $(y^*, \langle y^*, x\rangle) \in \operatorname{epi}\sigma_C$ by (2.4). Hence, if $\{C_i : i \in I\}$ satisfies the SECQ, one can apply (2.7) to express $(y^*, \langle y^*, x\rangle)$ as

$$(y^*, \langle y^*, x\rangle) = \sum_{j\in J}(y_j^*, u_j)$$

for some finite set $J \subseteq I$ and $(y_j^*, u_j) \in \operatorname{epi}\sigma_{C_j}(x)$ for each $j \in J$. Then $\langle y_j^*, x\rangle \le \sigma_{C_j}(y_j^*) \le u_j$ for all $j \in J$ and $\sum_{j\in J}\langle y_j^*, x\rangle = \sum_{j\in J} u_j$. It follows that $\langle y_j^*, x\rangle = u_j$ for each $j \in J$ and hence that $y_j^* \in N_{C_j}(x)$ by (2.4). Therefore $y^* \in \sum_{i\in I} N_{C_i}(x)$. Thus the strong CHIP for $\{C_i : i \in I\}$ is proved.

Conversely, assume that $\operatorname{dom}\sigma_C \subseteq \operatorname{Im}\partial\,\delta_C$ and that the strong CHIP for $\{C_i : i \in I\}$ is satisfied. We have to show that

$$\operatorname{epi}\sigma_C \subseteq \sum_{i\in I}\operatorname{epi}\sigma_{C_i}. \tag{3.2}$$

To do this, let $(y^*, \alpha) \in \operatorname{epi}\sigma_C$, that is, $\alpha \ge \sigma_C(y^*)$. Hence $y^* \in \operatorname{dom}\sigma_C$. Then, by the assumption and (2.4), there exists $x \in C$ such that $y^* \in N_C(x)$. By the strong CHIP assumption, it follows that there exist a finite index set $J \subseteq I$ and $y_j^* \in N_{C_j}(x)$ for each $j \in J$ such that

$$y^* = \sum_{j\in J} y_j^*. \tag{3.3}$$

Note that, for each $j \in J$, $\sigma_{C_j}(y_j^*) \le \langle y_j^*, x\rangle$ because $y_j^* \in N_{C_j}(x)$. Since $\alpha \ge \langle y^*, x\rangle = \sum_{j\in J}\langle y_j^*, x\rangle$, there exists a set $\{\alpha_j : j \in J\}$ of real numbers such that

$$\alpha = \sum_{j\in J}\alpha_j \quad\text{and}\quad \sigma_{C_j}(y_j^*) \le \langle y_j^*, x\rangle \le \alpha_j \quad \text{for each } j \in J.$$

This implies that $(y_j^*, \alpha_j) \in \operatorname{epi}\sigma_{C_j}$ for each j and $(y^*, \alpha) \in \sum_{i\in I}\operatorname{epi}\sigma_{C_i}$ thanks to (3.3). Hence (3.2) is proved. □

Given a closed convex set C and a finite dimensional linear subspace Y containing C, recall from [2, section 2.4] and [14] that $C \subseteq Y$ is said to be continuous if $y^* \mapsto \sup\{\langle y^*, y\rangle : y \in C\}$ is continuous on $Y^*\backslash\{0\}$. Here, the continuity at y_0^* with $\sup\{\langle y_0^*, y\rangle : y \in C\} = +\infty$ means that for each $\alpha \in \mathbb{R}$ there exists a neighborhood V_0 of y_0^* such that $\sup\{\langle v^*, y\rangle : y \in C\} > \alpha$ for all $v^* \in V_0$.

For convenience, we use $x^*|_Z$ to denote the restriction to Z of the functional $x^* \in X^*$, where Z is a linear subspace of X.

PROPOSITION 3.2. *Let C be a nonempty closed convex set in X. Then condition* (3.1) *holds in each of the following cases:*

(i) *There exists a weakly compact convex set D and a closed convex cone K such that $C = D + K$.*

(ii) $\dim C < \infty$, $\operatorname{Im}\partial\delta_C$ *is convex, and C is a continuous set as a subset of* $\operatorname{span} C$.

Proof. (i) Suppose that (i) holds and let $y^* \in \operatorname{dom}\sigma_C$. Then since K is a cone,

$$\sup_{d\in D}\langle y^*, d\rangle = \sup_{d\in D}\langle y^*, d\rangle + \sup_{k\in K}\langle y^*, k\rangle = \sup_{d\in D,\, k\in K}\langle y^*, d+k\rangle = \sigma_C(y^*) < +\infty. \tag{3.4}$$

Since D is weakly compact, there exists $\overline{x} \in D\ (\subseteq C)$ such that $\langle y^*, \overline{x}\rangle = \sup_{d\in D}\langle y^*, d\rangle$. Thus by (3.4), $\langle y^*, \overline{x}\rangle = \sigma_D(y^*) = \sigma_C(y^*)$. Hence $y^* \in N_C(\overline{x})$ and (3.1) is proved.

(ii) Suppose that (ii) holds. If C is bounded, then C is compact because $\operatorname{span} C$ is finite dimensional. Hence (3.1) in this case follows from part (i). If C is the whole space, then (3.1) holds trivially as $\operatorname{dom}\sigma_C = \operatorname{Im}\partial\,\delta_C = \{0\}$. Thus we may assume that C is a proper and unbounded subset of the finite dimensional space $Z := \operatorname{span} C$. Let $\hat{\delta}_C$ denote the indicator function of the set C as a set in the space Z, and let $\hat{\sigma}_C(c^*) := \sup\{\langle c^*, c\rangle : c \in C\}$ for each $c^* \in Z^*$; that is, $\hat{\delta}_C$ and $\hat{\sigma}_C$ are the indicator function and the support function of C as a subset of Z, respectively. It is easy to see from definitions that

$$\operatorname{dom}\sigma_C = \{y^* \in X^* : y^*|_Z \in \operatorname{dom}\hat{\sigma}_C\} \quad \text{and} \quad \operatorname{Im}\partial\delta_C = \{y^* \in X^* : y^*|_Z \in \operatorname{Im}\partial\,\hat{\delta}_C\}. \tag{3.5}$$

Now, by the assumption, it follows that $\operatorname{Im}\partial\,\hat{\delta}_C$ is convex in Z^*. We claim that

$$\operatorname{dom}\hat{\sigma}_C \subseteq \operatorname{Im}\partial\,\hat{\delta}_C. \tag{3.6}$$

Since C is proper, unbounded, and continuous as a subset of Z, we know from [2, Proposition 2.4.3] that

$$\operatorname{dom}\hat{\sigma}_C\backslash\{\mathbf{0}\} = \operatorname{int}(\operatorname{dom}\hat{\sigma}_C) \neq \emptyset. \tag{3.7}$$

On the other hand, since $\operatorname{Im}\partial\,\hat{\delta}_C$ is a convex set in the finite dimensional Banach space Z^*, one has (cf. [35, Proposition 1.2.1 and Corollary 1.3.4])

$$\operatorname{int}(\operatorname{Im}\partial\,\hat{\delta}_C) = \operatorname{int}(\overline{\operatorname{Im}\partial\,\hat{\delta}_C}). \tag{3.8}$$

Moreover, by [35, Theorem 3.1.2], one has $\operatorname{dom}\hat{\sigma}_C \subseteq \overline{\operatorname{Im}\partial\hat{\delta}_C}$. Consequently, by (3.6)–(3.8), we get that

$$\operatorname{dom}\hat{\sigma}_C\backslash\{\mathbf{0}\} = \operatorname{int}(\operatorname{dom}\hat{\sigma}_C) \subseteq \operatorname{int}(\overline{\operatorname{Im}\partial\,\hat{\delta}_C}) = \operatorname{int}(\operatorname{Im}\partial\,\hat{\delta}_C) \subseteq \operatorname{Im}\partial\,\hat{\delta}_C.$$

Therefore claim (3.6) stands because $0 \in \operatorname{Im}\partial\hat{\delta}_C$. Consequently, (3.1) follows from (3.5), (3.6), and the Hahn–Banach theorem. The proof is complete. □

Combining Theorem 3.1 and Proposition 3.2, we immediately have the following corollary.

COROLLARY 3.3. *Let $\{C_i : i \in I\}$ be a family of closed convex sets in X. Then the strong CHIP and the SECQ are equivalent for $\{C_i : i \in I\}$ in each of the following cases.*

(i) *There exists a weakly compact convex set D and a closed convex cone K such that $C = D + K$.*
(ii) $\dim C < \infty$, *$\operatorname{Im}\partial\delta_C$ is convex, and C is a continuous set as a subset of* $\operatorname{span} C$.

Remark 3.1. Part (i) was known in some special cases; see [7, Proposition 4.2] for the case when I is a two point set and $D = \{0\}$, and see [20] for the case when I is a finite set and $D = \{0\}$.

4. Linear regularity and the SECQ. Let I be an arbitrary index set and let $\{C_i : i \in I\}$ be a *CCS*-system with $0 \in C$, where $C = \bigcap_{i\in I} C_i$ as before. Throughout this section, we shall use $\mathbf{\Sigma}^*$ to denote the set $\mathbf{B}^* \times \mathbb{R}^+$, where $\mathbf{B}^*$ is the closed unit ball of X^*, while $\mathbb{R}^+$ consists of all nonnegative real numbers. This section is devoted to a study of the relationship between the linear regularity and the SECQ. We begin with the notion of the linear regularity for the system $\{C_i : i \in I\}$ and two simple lemmas (the first one is easy to verify). Recall that, for a closed convex set S in a normed linear space X, $d_S(\cdot)$ denotes the distance function of S.

DEFINITION 4.1. *The system $\{C_i : i \in I\}$ is said to be*

(i) *linearly regular if there exists a constant $\gamma > 0$ such that*

$$d_C(x) \le \gamma \sup_{i\in I} d_{C_i}(x) \quad \textit{for all } x \in X;$$

(ii) *boundedly linearly regular if, for each $r > 0$, there exists a constant $\gamma_r > 0$ such that*

$$d_C(x) \le \gamma_r \sup_{i\in I} d_{C_i}(x) \quad \textit{for all } x \in r\mathbf{B}.$$

LEMMA 4.2. *Let $\gamma > 0$. Then*

$$\overline{\operatorname{co} \cup_{i\in I}(\operatorname{epi} \sigma_{C_i} \cap \gamma\mathbf{\Sigma}^*)}^{w^*} + \{0\} \times \mathbb{R}^+ = \overline{\operatorname{co} \cup_{i\in I}(\operatorname{epi} \sigma_{C_i} \cap \gamma\mathbf{\Sigma}^*)}^{w^*}.$$

LEMMA 4.3. *Let $\gamma > 0$ and let $f_\gamma := \gamma d_S$. If $0 \in S$, then*

$$\operatorname{epi} f_\gamma^* = \operatorname{epi} \sigma_S \cap \left(\gamma\mathbf{B}^* \times \mathbb{R}^+\right). \tag{4.1}$$

Proof. By conjugation computation rules (cf. [35, Theorem 2.3.1(v) and Proposition 3.8.3(i)]), we have for any $x^* \in X^*$,

$$f_\gamma^*(x^*) = \gamma(\sigma_S + \delta_{\mathbf{B}^*})\left(\frac{x^*}{\gamma}\right) = \sigma_S(x^*) + \delta_{\gamma\mathbf{B}^*}(x^*).$$

Then (4.1) follows immediately. □

In the next two theorems, we shall use the graph $\operatorname{gph} f$ of a function f which is defined by

$$\operatorname{gph} f := \{(x, f(x)) \in X \times \mathbb{R} : \ x \in \operatorname{dom} f\}.$$

Clearly, $\operatorname{gph} f \subseteq \operatorname{epi} f$ for a function f on X.

THEOREM 4.4. *Let $\gamma > 0$. Then the following conditions are equivalent:*

(i) *For all $x \in X$, $d_C(x) \le \gamma \sup_{i\in I} d_{C_i}(x)$.*

(ii) $\operatorname{epi} \sigma_C \cap \mathbf{\Sigma}^* \subseteq \overline{\operatorname{co} \cup_{i\in I}(\operatorname{epi} \sigma_{C_i} \cap \gamma\mathbf{\Sigma}^*)}^{w^*}$.

(iii) $\operatorname{gph} \sigma_C \cap \mathbf{\Sigma}^* \subseteq \overline{\operatorname{co} \cup_{i\in I}(\operatorname{epi} \sigma_{C_i} \cap \gamma\mathbf{\Sigma}^*)}^{w^*}$.

Proof. By Lemmas 2.3 and 4.3, one has that

$$\operatorname{epi}\left(\sup_{i\in I} d_{C_i}\right)^* = \overline{\operatorname{co} \bigcup_{i\in I} \operatorname{epi} d_{C_i}^*}^{w^*} = \overline{\operatorname{co} \bigcup_{i\in I} (\operatorname{epi} \sigma_{C_i} \cap \mathbf{\Sigma}^*)}^{w^*}.$$

Noting that $\operatorname{epi} \sigma_S$ is a cone, we see that the equivalence of (i) and (ii) follows from (2.1) and Lemma 4.3.

By Lemma 4.2, (iii) implies that

$$\begin{aligned}\operatorname{epi}\sigma_C \cap \mathbf{\Sigma}^* &\subseteq \operatorname{gph}\sigma_C \cap \mathbf{\Sigma}^* + \{0\}\times\mathbb{R}^+\\ &\subseteq \overline{\operatorname{co}\cup_{i\in I}(\operatorname{epi}\sigma_{C_i}\cap\gamma\mathbf{\Sigma}^*)}^{w^*} + \{0\}\times\mathbb{R}^+\\ &= \overline{\operatorname{co}\cup_{i\in I}(\operatorname{epi}\sigma_{C_i}\cap\gamma\mathbf{\Sigma}^*)}^{w^*}.\end{aligned}$$

Therefore (iii)⇒(ii). Since (ii)⇒(iii) is obvious, the proof is complete. □

We give a simple application of our new characterization of the linear regularity in Theorem 4.4. The following theorem includes an important characterization of the linear regularity of finitely many closed convex sets in a Banach space, given in [30, Theorem 4.2].

THEOREM 4.5. *Let $\gamma > 0$ and suppose that X is a Banach space. Consider the following statements:*

(i) *For all $x \in X$, $d_C(x) \le \gamma \sup_{i\in I} d_{C_i}(x)$.*

(ii) *For all $x \in C$, $N_C(x)\cap\mathbf{B}^* \subseteq \overline{\operatorname{co}\cup_{i\in I}(N_{C_i}(x)\cap\gamma\mathbf{B}^*)}^{w^*}$.*

Then (ii) *implies* (i). *If we assume further that I is a compact metric space and $i \mapsto C_i$ is lower semicontinuous, then* (i) *and* (ii) *are equivalent. In particular, when I is finite,* (i), (ii), (ĩi), *and* (iii) *are equivalent, where* (ĩi) *and* (iii) *are defined in the following:*

(ĩi) *For all $x \in C$, $N_C(x)\cap\mathbf{B}^* \subseteq \operatorname{co}\cup_{i\in I}(N_{C_i}(x)\cap\gamma\mathbf{B}^*)$.*

(iii) *For all $x \in C$ and for all $x^* \in N_C(x)\cap\mathbf{B}^*$, there exist $x_i^* \in N_{C_i}(x)$, $i \in I$, such that $\sum_{i\in I}\|x_i^*\| \le \gamma$ and $x^* = \sum_{i\in I} x_i^*$.*

Remark 4.1. Let $\rho > 0$ and recall from [3, 30] that the collection $\{D_1,\dots,D_m\}$ in X is said to have property (G_ρ) if

$$\left(\sum_{i=1}^m D_i\right)\bigcap\mathbf{B} \subseteq \sum_{i=1}^m\left(D_i\bigcap\frac{1}{\rho}\mathbf{B}\right).$$

Clearly, when I is finite, there exists $\gamma > 0$ such that condition (iii) above holds if and only if the strong CHIP holds for all $x \in C$ and there exists $\rho > 0$ such that, for each $x \in C$, $\{N_{C_i}(x) : i \in I\}$ has the property (G_ρ) in X^*.

Proof. (ii)⇒(i). In view of Theorem 4.4, to establish (i), it is sufficient to show that

$$\operatorname{gph}\sigma_C \cap \mathbf{\Sigma}^* \subseteq \overline{\operatorname{co}\bigcup_{i\in I}(\operatorname{epi}\sigma_{C_i}\cap\gamma\mathbf{\Sigma}^*)}^{w^*}.$$

To do this, let $(y^*,\sigma_C(y^*)) \in \operatorname{gph}\sigma_C\cap\mathbf{\Sigma}^*$. We have to show that

$$(y^*,\sigma_C(y^*)) \in \overline{\operatorname{co}\bigcup_{i\in I}(\operatorname{epi}\sigma_{C_i}\cap\gamma\mathbf{\Sigma}^*)}^{w^*}. \tag{4.2}$$

Since the set on the right-hand side of (4.2) obviously contains the origin, we may suppose without loss of generality that $y^* \ne 0$.

Consider first the case when $y^* \in \operatorname{Im}\partial\,\delta_C$. Then $y^* \in N_C(x)\cap\mathbf{B}^*$ for some $x \in C$ by (2.3). Thus one can apply (ii) to find a net $\{\tilde{y}_V^*\}$ with w^*-limit y^* such that for each V, $\tilde{y}_V^*$ is representable as

$$\tilde{y}_V^* = \gamma\sum_{i\in J_V}\lambda_i y_i^*$$

for some finite index set $J_V \subseteq I$, $y_i^* \in N_{C_i}(x) \cap \mathbf{B}^*$, $i \in J_V$, and $\lambda_i \in [0,1]$ with $\sum_{i\in J_V} \lambda_i = 1$. Using (2.4) again, we obtain $(y_i^*, \langle y_i^*, x\rangle) \in \operatorname{epi} \sigma_{C_i}$ for each $i \in J_V$. In w^*-limits, it follows that

$$(y^*, \langle y^*, x\rangle) = \lim_V (\tilde{y}_V^*, \langle \tilde{y}_V^*, x\rangle) = \lim_V \gamma \sum_{i\in J_V} \lambda_i (y_i^*, \langle y_i^*, x\rangle);$$

hence,

$$(y^*, \langle y^*, x\rangle) \in \overline{\operatorname{co} \bigcup_{i\in I} (\operatorname{epi} \sigma_{C_i} \cap \gamma \mathbf{\Sigma}^*)}^{w^*} \tag{4.3}$$

and, in particular, (4.2) holds provided that $y^* \in \operatorname{Im} \partial\, \delta_C$ and $\|y^*\| \le 1$. For the general case (that is, we do not assume that $y^* \in \operatorname{Im} \partial\, \delta_C$), by [35, Theorem 3.1.4(ii)], there exists a sequence $(y_n, y_n^*) \in \operatorname{gph} \partial\, \delta_C$ such that y_n^* converges to y^* in norm and $\sigma_C(y_n^*)$ converges to $\sigma_C(y^*)$. Note that by (2.4), we have $(y_n^*, \langle y_n^*, y_n\rangle) \in \operatorname{gph} \sigma_C$. If $\|y_n^*\| \le 1$ for all but finitely many $n \in \mathbb{N}$, then one can apply (4.3) to $(y_n^*, \sigma_C(y_n^*))$ in place of $(y^*, \sigma_C(y^*))$ to conclude that

$$(y_n^*, \sigma_C(y_n^*)) \in \overline{\operatorname{co} \bigcup_{i\in I} (\operatorname{epi} \sigma_{C_i} \cap \gamma \mathbf{\Sigma}^*)}^{w^*}. \tag{4.4}$$

On the other hand, if it happens that $\|y_n^*\| > 1$ for infinitely many $n \in \mathbb{N}$, then we must have $\|y^*\| = 1$ and $\|y_n^*\| \to 1$ as $n \to \infty$. Since $\operatorname{Im} \partial\, \delta_C$ is a cone, we see that $\frac{y_n^*}{\|y_n^*\|} \in \operatorname{Im} \partial\, \delta_C$. Applying (4.3) to $(\frac{y_n^*}{\|y_n^*\|}, \sigma_C(\frac{y_n^*}{\|y_n^*\|}))$ in place of $(y^*, \sigma_C(y^*))$, we obtain

$$\left(\frac{y_n^*}{\|y_n^*\|}, \sigma_C\left(\frac{y_n^*}{\|y_n^*\|}\right)\right) \in \overline{\operatorname{co} \bigcup_{i\in I} (\operatorname{epi} \sigma_{C_i} \cap \gamma \mathbf{\Sigma}^*)}^{w^*}. \tag{4.5}$$

Taking limits in (4.4) and (4.5), we get (4.2), as required in both cases. This completes the proof of (ii)⇒(i).

For the converse implication, let us begin by noting that if (i) holds, then by the definition of subdifferential, we have

$$\partial\, d_C(x) \subseteq \gamma\, \partial \sup_{i\in I} d_{C_i}(x) \quad \text{for each } x \in C. \tag{4.6}$$

Now we suppose further that I is compact and that $i \mapsto C_i$ is lower semicontinuous. Then by [1, Corollary 1.4.17], $i \mapsto d_{C_i}(\cdot)$ is upper semicontinuous. Hence one can apply [35, Theorem 2.4.18] to get the inclusion $\partial \left(\sup_{i\in I} d_{C_i}\right)(x) \subseteq \overline{\operatorname{co} \cup_{i\in I} \partial\, d_{C_i}(x)}^{w^*}$, and it follows from (4.6) that $\partial\, d_C(x) \subseteq \gamma\, \overline{\operatorname{co} \cup_{i\in I} \partial\, d_{C_i}(x)}^{w^*}$; hence (ii) holds for all $x \in C$ thanks to the standard result that $\partial\, d_C(x) = N_C(x) \cap B^*$ and $\partial\, d_{C_i}(x) = N_{C_i}(x) \cap B^*$ (cf. [35, Proposition 3.8.3]).

Next, we consider the case when I is finite. We need only show that (ii)⇔($\tilde{\text{ii}}$) in this case. For any $x \in C$, we note that by the Banach–Alaoglu theorem, $N_{C_i}(x) \cap \mathbf{B}^*$ is w^*-compact for each $i \in I$; thus $\operatorname{co} \cup_{i\in I} (N_{C_i}(x) \cap \mathbf{B}^*)$ is w^*-closed as I is finite. Hence (ii) and ($\tilde{\text{ii}}$) are the same when I is finite.

Finally, we turn to prove that ($\tilde{\text{ii}}$)⇔(iii). The forward implication is obvious. For the converse implication, fix $x \in C$. Let $x^* \in N_C(x) \cap \mathbf{B}^*$; we wish to show that $x^* \in \operatorname{co} \cup_{i\in I} (N_{C_i}(x) \cap \gamma \mathbf{B}^*)$. By (iii), there exist $x_i^* \in N_{C_i}(x)$, $i \in I$, with

$\sum_{i\in I}\|x_i^*\| \le \gamma$ and $x^* = \sum_{i\in I} x_i^*$. If all the x_i^*'s are zero, then the inclusion holds trivially. Otherwise, set $\lambda := \sum_{i\in I}\|x_i^*\| > 0$. Then $\lambda \le \gamma$. Thus we see that

$$x^* = \lambda\left(\sum_{i\in I,\, x_i^*\neq 0}\frac{\|x_i^*\|}{\lambda}\frac{x_i^*}{\|x_i^*\|} + \left(1 - \sum_{i\in I,\, x_i^*\neq 0}\frac{\|x_i^*\|}{\lambda}\right)0\right) \in \operatorname{co}\bigcup_{i\in I}(N_{C_i}(x)\cap\gamma\mathbf{B}^*),$$

which completes the proof. □

THEOREM 4.6. *Suppose that*

$$\overline{\operatorname{co}\bigcup_{i\in I}(\operatorname{epi}\sigma_{C_i}\cap\mathbf{\Sigma}^*)}^{w^*} \subseteq \sum_{i\in I}\operatorname{epi}\sigma_{C_i}, \tag{4.7}$$

and that $\{C_i : i\in I\}$ is linearly regular. Then it satisfies the SECQ.

Proof. By the assumption, one can combine (4.7) with Theorem 4.4 to conclude that $\operatorname{epi}\sigma_C\cap\mathbf{\Sigma}^* \subseteq \sum_{i\in I}\operatorname{epi}\sigma_{C_i}$, and hence that $\operatorname{epi}\sigma_C \subseteq \sum_{i\in I}\operatorname{epi}\sigma_{C_i}$ for each $\operatorname{epi}\sigma_{C_i}$ is a cone. □

In the next theorem, we shall provide some sufficient conditions for (4.7). We first prove a simple lemma. We shall prove it in a bit more general context for later use. Recall that $\{C_i : i\in I\}$ is a *CCS*-system with $0\in C$.

LEMMA 4.7. *Let I be a metric space. Suppose that Z is a linear subspace of X and $i\mapsto Z\cap C_i$ is lower semicontinuous. Consider elements $i_0\in I$, $(x_0^*,\alpha_0)\in X^*\times\mathbb{R}$ and nets $\{i_k\}\subseteq I$, $\{(x_k^*,\alpha_k)\}\subseteq X^*\times\mathbb{R}$ with each $(x_k^*,\alpha_k)\in\operatorname{epi}\sigma_{C_{i_k}}$. Suppose further that $i_k\to i_0$, $\alpha_k\to\alpha_0$, and $x_k^*|_Z \to^{w^*} x_0^*|_Z$. If $\{x_k^*|_Z\}$ is bounded, then $(x_0^*,\alpha_0)\in\operatorname{epi}\sigma_{Z\cap C_{i_0}}$.*

Proof. Let $x\in Z\cap C_{i_0}$. We have to prove that $\langle x_0^*, x\rangle \le \alpha_0$. By the assumption, there exists a net $\{x_k\}\subseteq X$ with each $x_k\in Z\cap C_{i_k}$ such that $x_k\to x$. Since

$$\langle x_0^*, x\rangle = \langle x_0^* - x_k^*, x\rangle + \langle x_k^*, x - x_k\rangle + \langle x_k^*, x_k\rangle,$$

where on the right-hand side the first two terms converge to zero and the last term $\langle x_k^*, x_k\rangle \le \alpha_k$ for each k, it follows by passing to the limits that $\langle x_0^*, x\rangle \le \alpha_0$. □

THEOREM 4.8. *Let I be a compact metric space and $i\mapsto C_i$ be lower semicontinuous on I. Suppose that either I is finite or there exists an index $i_0\in I$ such that $\dim C_{i_0} < +\infty$. Then (4.7) holds. Consequently, if $\{C_i : i\in I\}$ is, in addition, linearly regular, then it satisfies the SECQ.*

Proof. We first assume that I is finite, say $I = \{1,2,\ldots,m\}$. Let $(\overline{x}^*,\overline{\alpha})\in \overline{\operatorname{co}\cup_{i=1}^m(\operatorname{epi}\sigma_{C_i}\cap\mathbf{\Sigma}^*)}^{w^*}$. Then there exists a net $\{(\overline{x}_k^*,\overline{\alpha}_k)\}$ in $\operatorname{co}\cup_{i=1}^m(\operatorname{epi}\sigma_{C_i}\cap\mathbf{\Sigma}^*)$ such that $(\overline{x}_k^*,\overline{\alpha}_k)$ w^*-converges to $(\overline{x}^*,\overline{\alpha})$. Without loss of generality, we assume that $0\le\overline{\alpha}_k\le\overline{\alpha}+1$ for all k. Each $(\overline{x}_k^*,\overline{\alpha}_k)$ can be expressed as a convex combination

$$(\overline{x}_k^*,\overline{\alpha}_k) = \sum_{i=1}^m \lambda_{k,i}(x_{k,i}^*,\alpha_{k,i}) \tag{4.8}$$

for some $(x_{k,i}^*,\alpha_{k,i})\in\operatorname{epi}\sigma_{C_i}\cap\Sigma^*$ and $\lambda_{k,i}\in[0,1]$ with $\sum_{i=1}^m\lambda_{k,i}=1$. Note that

$$\lambda_{k,i}(x_{k,i}^*,\alpha_{k,i})\in\operatorname{epi}\sigma_{C_i}\cap\mathbf{\Sigma}^* \quad\text{for each } k \text{ and } i. \tag{4.9}$$

By considering subnets if necessary and by the w^*-compactness of the closed unit ball in Banach dual space X^* (the Banach–Alaoglu theorem), we may assume without loss of generality that for each i, there exist $x_i^*\in\mathbf{B}^*$ and $\beta_i\in[0,\overline{\alpha}+1]$ such that

$$\lambda_{k,i}x_{k,i}^*\to x_i^*, \quad \lambda_{k,i}\alpha_{k,i}\to\beta_i \tag{4.10}$$

(note that $\lambda_{k,i}\alpha_{k,i} \le \overline{\alpha}+1$ for all k). By the w^*-closedness of the set $\operatorname{epi}\sigma_{C_i}$, we have from (4.10) and (4.9) that

$$(x_i^*, \beta_i) \in \operatorname{epi}\sigma_{C_i} \quad \text{for each } i. \tag{4.11}$$

Passing to the limits in (4.8), we arrive at

$$(\overline{x}^*, \overline{\alpha}) = \sum_{i=1}^{m}(x_i^*, \beta_i) \in \sum_{i=1}^{m}\operatorname{epi}\sigma_{C_i},$$

where the inclusion follows from (4.11).

Next we assume that there exists an index $i_0 \in I$ such that $\dim C_{i_0} < +\infty$. Let $Z_0 = \operatorname{span} C_{i_0}$ and let $(\overline{x}^*, \overline{\alpha}) \in \overline{\operatorname{co}\cup_{i\in I}(\operatorname{epi}\sigma_{C_i} \cap \mathbf{\Sigma}^*)}^{w^*}$. Then there exists a net $\{(\overline{x}_k^*, \overline{\alpha}_k)\}$ in $\operatorname{co}\cup_{i\in I}(\operatorname{epi}\sigma_{C_i} \cap \mathbf{\Sigma}^*)$ such that $(\overline{x}_k^*, \overline{\alpha}_k) \to^{w^*} (\overline{x}^*, \overline{\alpha})$. Since $Z_0 \times \mathbb{R}$ is of dimension $m+1$, one can apply the Carathéodory theorem to express each $(\overline{x}_k^*, \overline{\alpha}_k)$ as a convex combination of $m+2$ many elements of $\cup_{i\in I}(\operatorname{epi}\sigma_{C_i} \cap \mathbf{\Sigma}^*)$. Hence there exist indices $i_j^k \in I$, nonnegative scalars $\lambda_{k,j}$, and pairs

$$(x_{k,j}^*, \alpha_{k,j}) \in \operatorname{epi}\sigma_{C_{i_j^k}} \cap \mathbf{\Sigma}^* \quad \text{for each } 1 \le j \le m+2$$

with the properties $\sum_{j=1}^{m+2}\lambda_{k,j} = 1$ and

$$(\overline{x}_k^*|_{Z_0}, \overline{\alpha}_k) = \sum_{j=1}^{m+2}\lambda_{k,j}(x_{k,j}^*|_{Z_0}, \alpha_{k,j}). \tag{4.12}$$

Note that

$$\lambda_{k,j}(x_{k,j}^*, \alpha_{k,j}) \in \operatorname{epi}\sigma_{C_{i_j^k}} \cap \mathbf{\Sigma}^*. \tag{4.13}$$

Since $\{\overline{\alpha}_k\}$ is convergent, by passing to subnets if necessary, we may assume that $\overline{\alpha}+1 \ge \overline{\alpha}_k \ge 0$. Then we also have $\{\overline{\alpha}_k\}$ and $\{\lambda_{k,j}\alpha_{k,j}\}$ bounded for $1 \le j \le m+2$. Hence, considering subnets if necessary, we may assume that each of the nets $\{\lambda_{k,j}x_{k,j}^*\}$, $\{\overline{\alpha}_k\}$, $\{\lambda_{k,j}\alpha_{k,j}\}$ for $1 \le j \le m+2$ converges, say with limits,

$$x_{0,j}^*, \quad \overline{\alpha}, \quad \alpha_{0,j},$$

and we can assume further that i_j^k converges to some $i_j^0 \in I$ $(1 \le j \le m+2)$, thanks to the compactness assumption of I. Making use of (4.13) and thanks to the assumption that $i \mapsto C_i$ is lower semicontinuous, it follows from Lemma 4.7 (applied to X in place of Z) that

$$(x_{0,j}^*, \alpha_{0,j}) \in \operatorname{epi}\sigma_{C_{i_j^0}} \quad \text{for each } 1 \le j \le m+2.$$

Moreover, passing to the limits in (4.12), we have

$$(\overline{x}^*|_{Z_0}, \overline{\alpha}) = \sum_{j=1}^{m+2}(x_{0,j}^*|_{Z_0}, \alpha_{0,j}).$$

Noting the trivial relation that $\operatorname{epi}\sigma_{C_{i_0}}$ contains $Z_0^\perp \times \mathbb{R}^+$, where $Z_0^\perp := \{x^* \in X^* : x^*|_{Z_0} = 0\}$, it follows that

$$(\overline{x}^*, \overline{\alpha}) \in \sum_{j=1}^{m+2}(x_{0,j}^*, \alpha_{0,j}) + Z_0^\perp \times \mathbb{R}^+ \subseteq \sum_{i\in I}\operatorname{epi}\sigma_{C_i}.$$

This shows that (4.7) holds.

Finally, in addition we assume that $\{C_i : i \in I\}$ is linearly regular. Then it follows from Theorem 4.6 that this system satisfies the SECQ. □

We intend to relate bounded linear regularity with the strong CHIP. We first provide a sufficient condition for a system to be linearly regular. The result is known when the ambient space is a Hilbert space [4, Theorem 4.2.6, Corollary 4.4.4] or a Banach space [34, Corollary 5]. The corresponding theorems in those references are derived from a lemma whose proof is based on the open mapping theorem and thus does not work in general normed linear spaces. As some preparatory work, we first state the following lemma, which is a generalization of [7, Proposition 3.1(i)] to a normed linear space (or even locally convex space) setting. We shall omit its proof, as it is a direct application of [29, Remarque 10.2] (alternatively, the proof given for [7, Proposition 3.1(i)], which was based on a result in [34], can easily be adopted here).

LEMMA 4.9. *Let E, F be two closed convex sets in X with $E \cap \operatorname{int} F \neq \emptyset$. Then $\{E, F\}$ satisfies the SECQ.*

We now give a sufficient condition for a system to be linearly regular.

LEMMA 4.10. *Let E be a closed convex set in X containing the origin and let $r > 0$. Then*

$$d_{E\cap r\mathbf{B}}(x) \leq 4\max\{d_E(x), d_{r\mathbf{B}}(x)\} \quad \text{for each } x \in X. \tag{4.14}$$

Proof. We first show that

$$\operatorname{gph}\sigma_{E\cap r\mathbf{B}} \cap \mathbf{\Sigma}^* \subseteq \operatorname{co}\left((\operatorname{epi}\sigma_E \cap 4\mathbf{\Sigma}^*) \cup (\operatorname{epi}\sigma_{r\mathbf{B}} \cap 4\mathbf{\Sigma}^*)\right). \tag{4.15}$$

Take $(y^*, \sigma_{E\cap r\mathbf{B}}(y^*)) \in \operatorname{gph}\sigma_{E\cap r\mathbf{B}} \cap \mathbf{\Sigma}^*$. By Lemma 4.9, there exist $(y_1^*, \alpha_1) \in \operatorname{epi}\sigma_E$ and $(y_2^*, \alpha_2) \in \operatorname{epi}\sigma_{r\mathbf{B}}$ such that

$$(y^*, \sigma_{E\cap r\mathbf{B}}(y^*)) = (y_1^*, \alpha_1) + (y_2^*, \alpha_2).$$

This implies that

$$\sigma_{E\cap r\mathbf{B}}(y^*) = \alpha_1 + \alpha_2. \tag{4.16}$$

Since $0 \in E$, we have $0 \leq \sigma_E(y_1^*) \leq \alpha_1$ and hence $\alpha_2 \leq \sigma_{E\cap r\mathbf{B}}(y^*) \leq r$ thanks to (4.16). It follows that $r\|y_2^*\| = \sigma_{r\mathbf{B}}(y_2^*) \leq \alpha_2 \leq r$, and thus $\|y_1^*\| \leq \|y^*\| + \|y_2^*\| \leq 2$. Therefore,

$$(y^*, \sigma_{E\cap r\mathbf{B}}(y^*)) = \frac{1}{2}\left[(2y_1^*, 2\alpha_1) + (2y_2^*, 2\alpha_2)\right] \in \operatorname{co}((\operatorname{epi}\sigma_E \cap 4\mathbf{\Sigma}^*) \cup (\operatorname{epi}\sigma_{r\mathbf{B}} \cap 4\mathbf{\Sigma}^*))$$

and (4.15) is established. By the implication (iii)⇒(i) of Theorem 4.4 (with $\gamma = 4$), it follows that (4.14) holds. □

The following proposition on a relationship between bounded linear regularity and the linear regularity was shown in [4, Theorem 4.2.6(ii)] for the special case when X is a Hilbert space.

PROPOSITION 4.11. *Let $\{A_i : i \in I\}$ be a system of closed convex sets in X containing the origin, and suppose that $\{A_i : i \in I\}$ is boundedly linearly regular. Then for all $r > 0$, the system $\{r\mathbf{B}, A_i : i \in I\}$ is linearly regular.*

Proof. Write $A = \cap_{i\in I} A_i$ and let $r > 0$. By assumption, there exists $k_r > 0$ such that

$$d_A(x) \leq k_r \sup_{i\in I} d_{A_i}(x) \quad \text{for each } x \in r\mathbf{B}. \tag{4.17}$$

Let f be defined by $f(x) := k_r \sup_{i\in I} d_{A_i}(x) - d_A(x)$ for each $x \in X$. From (4.17), we see that $f(x) \geq 0$ for all $x \in r\mathbf{B}$, and the equality holds for all $x \in \bigcap_{i\in I} A_i \cap r\mathbf{B}$. Since f is clearly Lipschitz with modulus $k_r + 1$, it follows from [8, Proposition 2.4.3] that $f(x) + (k_r + 1)d_{r\mathbf{B}}(x) \geq 0$ for all $x \in X$. This implies

$$d_A(x) \leq (2k_r + 1) \max\left\{d_{r\mathbf{B}}(x), \sup_{i\in I} d_{A_i}(x)\right\} \quad \text{for each } x \in X.$$

It follows from Lemma 4.10 that

$$\begin{aligned} d_{A\cap r\mathbf{B}}(x) &\leq 4\max\{d_{r\mathbf{B}}(x), d_A(x)\} \\ &\leq 4(2k_r + 1) \max\left\{d_{r\mathbf{B}}(x), \sup_{i\in I} d_{A_i}(x)\right\} \quad \text{for each } x \in X. \end{aligned}$$

This completes the proof. □

For the following corollary, we need a lemma, which will also be used in the next section.

LEMMA 4.12. *Let $\{D, C_i : i \in I\}$ be a family of closed convex sets with nonempty intersection. Let A be a closed subset of X such that*

$$D \bigcap \left(\bigcap_{i\in I} C_i\right) \bigcap \operatorname{int} A \neq \emptyset. \tag{4.18}$$

If $\{D, C_i : i \in I\}$ has the strong CHIP, then so does $\{D \cap A, C_i : i \in I\}$. As a partial converse result, if $\{D \cap A, C_i : i \in I\}$ has the strong CHIP at some point $a \in D \cap (\cap_{i\in I} C_i) \cap \operatorname{int} A$, so does $\{D, C_i : i \in I\}$.

Proof. Set $E := D \cap (\cap_{i\in I} C_i)$. By hypothesis, $N_E(x) = N_D(x) + \sum_{i\in I} N_{C_i}(x)$ for each $x \in E$. Since $E \cap \operatorname{int} A \neq \emptyset$ and $D \cap \operatorname{int} A \neq \emptyset$,

$$\begin{aligned} N_{(A\cap D)\cap(\cap_{i\in I} C_i)}(x) = N_{A\cap E}(x) &= N_A(x) + N_E(x) = N_A(x) + N_D(x) + \sum_{i\in I} N_{C_i}(x) \\ &= N_{A\cap D}(x) + \sum_{i\in I} N_{C_i}(x) \qquad \text{for each } x \in A \cap E. \end{aligned}$$

This proves the first part. For the second part, observe that $N_A(a) = \{0\}$; hence $N_{A\cap D}(a) = N_D(a)$ and $N_{A\cap E}(a) = N_E(a)$. The conclusion is then immediate. □

COROLLARY 4.13. *Suppose that I is a compact metric space and the set-valued function $i \mapsto C_i$ is lower semicontinuous. Suppose that either I is finite or there exists an index $i_0 \in I$ such that $\dim C_{i_0} < +\infty$. If $\{C_i : i \in I\}$ is boundedly linearly regular, then it has the strong CHIP.*

Proof. Fix any $x \in \cap_{i\in I} C_i$. Let $r = \|x\| + 1$. Since $\{C_i : i \in I\}$ is boundedly linearly regular, we obtain from Proposition 4.11 that $\{r\mathbf{B}, C_i : i \in I\}$ is linearly regular. Taking an index $i_\infty \notin I$, set $I_\infty = I \cup \{i_\infty\}$ and $C_{i_\infty} = r\mathbf{B}$. Clearly the map $i \mapsto C_i$ is lower semicontinuous on I_∞. It now follows from the assumptions and Theorem 4.8 that $\{C_i : i \in I_\infty\}$ satisfies the SECQ and so does $\{r\mathbf{B}, C_i : i \in I\}$; thus $\{r\mathbf{B}, C_i : i \in I\}$ has the strong CHIP (thanks to Theorem 3.1). Then it follows from Lemma 4.12 (with $D = X$ and $A = r\mathbf{B}$) that $\{C_i : i \in I\}$ has the strong CHIP at x because $x \in \operatorname{int}(r\mathbf{B})$. The proof is complete. □

Corollary 4.13 and the implication (i)$\Longrightarrow$(ii) in Theorem 4.5 were established under the following conditions: (α) I is compact, and (β) the set-valued function

$i \mapsto C_i$ is lower semicontinuous. Below we give a simple example with several cases to illustrate that these conditions are needed.

Example 4.1. Let $X = \mathbb{R}^2$ and $J = \{\frac{1}{n} : n \in \mathbb{N}\}$. Define $C_j := \{x \in X : \|x\| \le j\}$ for each $j \in J$. Consider the system $\{X, C_i : i \in I\}$ with I defined in the following:

(a) $I = J$.

(b) $I = \{0\} \cup J$ and $C_0 := \{x \in X : \|x\| \le 1\}$.

(c) $I = \{0\} \cup J$ and $C_0 := \{0\}$.

Then, in each case, $C = \cap_{i\in I} C_i = \{0\}$. Moreover, for each $x \in X$, $d_{C_i}(x) = \max\{0, \|x\| - i\}$ for each $i \in J$, while $d_{C_0}(x) = \max\{0, \|x\| - 1\}$ in (b) and $d_{C_0}(x) = \|x\|$ in (c). It follows that

$$\sup_{i\in I} d_{C_i}(x) = \|x\| = d(x, 0) = d_{\bigcap_{i\in I} C_i}(x). \tag{4.19}$$

Thus the system $\{C_i : i \in I\}$ is linearly regular in each of (a), (b), and (c). Note also that (c) satisfies both (α) and (β) and hence that Theorem 4.5(ii) and the conclusion of Corollary 4.13 hold (and these can be verified directly too). For each of (a) and (b), we have $C = \{0\}$ and $N_{C_i}(0) = \{0\}$ for each $i \in I$, so $N_C(0) \cap \mathbf{B} = \mathbf{B}$ but $\operatorname{co} \cup_{i\in I}(N_{C_i}(0) \cap \mathbf{B}) = \{0\}$; thus the corresponding system does not have the strong CHIP (nor the SECQ, by Theorem 3.1), and Theorem 4.5(ii) does not hold. This failure of (a) and (b) is because each satisfies only one condition of (α), (β) (I is not compact in (a) and the set-valued function $i \mapsto C_i$ is not lower semicontinuous in (b)).

5. Interior-point conditions and the SECQ. Recall that I is an index-set and $C = \bigcap_{i\in I} C_i \subseteq X$. As in [27], the family $\{D, C_i : i \in I\}$ is called a closed convex set system with base-set D (*CCS*-system with base-set D) if D and all C_i's are closed convex subsets of X. Furthermore, throughout the remainder of this section, we always assume that I is a compact metric space and that $0 \in D \cap C$. Thus,

$$\sigma_D \text{ and } \sigma_{C_i} \text{ are nonnegative functions on } X^* \text{ for all } i \in I.$$

Let $|J|$ denote the cardinality of the set J.

DEFINITION 5.1. *Let $\{D, C_i : i \in I\}$ be a CCS-system with base-set D. Let m be a positive integer. Then the CCS-system $\{D, C_i : i \in I\}$ is said to satisfy*

(i) *the m-D-interior-point condition if, for any subset J of I with $|J| \le m$,*

$$D \bigcap \left(\bigcap_{i\in J} \operatorname{rint}_D C_i \right) \neq \emptyset;$$

(ii) *the m-interior-point condition if, for any subset J of I with $|J| \le m$,*

$$D \bigcap \left(\bigcap_{i\in J} \operatorname{int} C_i \right) \neq \emptyset.$$

Before proving our main theorems, we first give the following lemma. Recall that, for a linear subspace Z of X, $y^*|_Z \in Z^*$ is the restriction to Z of y^*.

LEMMA 5.2. *Let m be a positive integer and let $\{D, C_i : i \in I\}$ be a CCS-system with the base-set D. Let $Z := \operatorname{span} D$ and suppose that the following conditions are satisfied:*

(a) *D is finite dimensional.*

(b) *The set-valued mapping $i \mapsto Z \cap C_i$ is lower semicontinuous on I.*
(c) *The system $\{D, C_i : i \in I\}$ satisfies the m-D-interior-point condition.*
Let $(y^*, \alpha) \in X^* \times \mathbb{R}$ and let $\{(y_k^*, \alpha_k)\} \subseteq X^* \times \mathbb{R}$ be a sequence such that

$$(y_k^*|_Z, \alpha_k) \text{ converges to } (y^*|_Z, \alpha), \tag{5.1}$$

where each $(y_k^*|_Z, \alpha_k)$ can be expressed in the form

$$(y_k^*|_Z, \alpha_k) = (v_k^*|_Z, \beta_k) + \sum_{j=1}^{m} \left(x_{i_j^k}^*\Big|_Z, \alpha_{i_j^k} \right) \tag{5.2}$$

with

$$(v_k^*, \beta_k) \in \operatorname{epi} \sigma_D, \quad \left(x_{i_j^k}^*, \alpha_{i_j^k} \right) \in \operatorname{epi} \sigma_{C_{i_j^k}} \tag{5.3}$$

for some $i_1^k, \ldots, i_m^k \in I$. Then

$$(y^*, \alpha) \in \operatorname{epi} \sigma_D + \sum_{i \in I} \operatorname{epi} \sigma_{Z \cap C_i}. \tag{5.4}$$

Proof. Since I is compact, by considering subsequences if necessary we may assume that there exists $i_j \in I$ such that $i_j^k \to i_j$ for each $j = 1, \ldots, m$. By assumption (c), there exist $z \in D$ and $\delta' > 0$ such that

$$\mathbf{B}(z, \delta') \cap Z \subseteq C_{i_j} \cap Z \quad \text{for each } j = 1, 2, \ldots, m. \tag{5.5}$$

Set for convenience $B := \mathbf{B}(z, \delta) \cap Z$, where $\delta = \frac{\delta'}{2}$. Then B is compact, thanks to assumption (a). For each $j = 1, 2, \ldots, m$, we make use of assumption (b) and apply Proposition 2.9 at the point $t_0 := i_j$ of the lower semicontinuous function $i \mapsto C_i \cap Z$ to conclude from (5.5) that $B \subseteq C_{i_j^k} \cap Z$ for all large enough k. Do this for each $j = 1, 2, \ldots, m$ and take $k_0 \in \mathbb{N}$ large enough such that

$$B \subseteq C_{i_j^k} \cap Z \quad \text{for each } 1 \le j \le m \text{ and } k \ge k_0. \tag{5.6}$$

Note that, for each $1 \le j \le m$ and $k \in \mathbb{N}$,

$$\sigma_B \left(x_{i_j^k}^* \right) = \sup_{x \in B} \left\langle x_{i_j^k}^*, x \right\rangle = \sup_{x \in \delta \mathbf{B} \cap Z} \left\langle x_{i_j^k}^*, x \right\rangle + \left\langle x_{i_j^k}^*, z \right\rangle = \delta \left\| x_{i_j^k}^*\Big|_Z \right\| + \left\langle x_{i_j^k}^*, z \right\rangle.$$

It follows from (5.6) that

$$\alpha_{i_j^k} \ge \sigma_{C_{i_j^k}} \left(x_{i_j^k}^* \right) \ge \sigma_{C_{i_j^k} \cap Z} \left(x_{i_j^k}^* \right) \ge \sigma_B \left(x_{i_j^k}^* \right) = \delta \left\| x_{i_j^k}^*\Big|_Z \right\| + \left\langle x_{i_j^k}^*, z \right\rangle, \tag{5.7}$$

provided that $k \ge k_0$. Moreover, since $z \in D$ and $(v_k^*, \beta_k) \in \operatorname{epi} \sigma_D$, (5.2) establishes that

$$\alpha_k - \langle y_k^*, z \rangle = \beta_k - \langle v_k^*, z \rangle + \sum_{j=1}^{m} \left(\alpha_{i_j^k} - \left\langle x_{i_j^k}^*, z \right\rangle \right) \ge \sum_{j=1}^{m} \left(\alpha_{i_j^k} - \left\langle x_{i_j^k}^*, z \right\rangle \right). \tag{5.8}$$

Combining (5.7) and (5.8) yields that

$$\alpha_k - \langle y_k^*, z \rangle \ge \sum_{j=1}^{m} \left(\alpha_{i_j^k} - \left\langle x_{i_j^k}^*, z \right\rangle \right) \ge \sum_{j=1}^{m} \delta \left\| x_{i_j^k}^*\Big|_Z \right\|.$$

This implies that $\{x^*_{i^k_j}|_Z : k \in \mathbb{N}\}$ is bounded for each $1 \le j \le m$ thanks to (5.1). Consequently $\{v^*_k|_Z : k \in \mathbb{N}\}$ is bounded as, by (5.2), $(v^*_k|_Z)$ is the sum of $m+1$ bounded sequences. Since Z is finite dimensional (and by passing to subsequences if necessary) we may assume that for each $j = 1, 2, \ldots, m$, there exist $\tilde{x}^*_{i_j}$ and $\tilde{v}^* \in Z^*$ such that

$$x^*_{i^k_j}\Big|_Z \to \tilde{x}^*_{i_j} \text{ and } v^*_k|_Z \to \tilde{v}^* \text{ as } k \to \infty.$$

Now, observe from (5.1)–(5.3) that $\{\alpha_{i^k_j}\}$ and $\{\beta_k\}$ are bounded. Thus we may also assume that, for each j, $\alpha_{i^k_j} \to \hat{\alpha}_{i_j}$ for some $\hat{\alpha}_{i_j} \in \mathbb{R}$ and that $\beta_k \to \hat{\beta}$ for some $\hat{\beta} \in \mathbb{R}$. Then, by (5.1) and (5.2),

$$y^*|_Z = \tilde{v}^* + \sum_{j=1}^m \tilde{x}^*_{i_j} \quad \text{and} \quad \alpha = \hat{\beta} + \sum_{j=1}^m \hat{\alpha}_{i_j}. \tag{5.9}$$

Let $x^*_{i_j} \in X^*$ be an extension of $\tilde{x}^*_{i_j}$ to X and $v^* \in X^*$ be an extension of $\tilde{v}^*$ to X. It follows from Lemma 4.7 that $(x^*_{i_j}, \hat{\alpha}_{i_j}) \in \operatorname{epi} \sigma_{(C_{i_j} \cap Z)}$ and $(v^*, \hat{\beta}) \in \operatorname{epi} \sigma_D$. Write $\hat{y}^* = y^* - v^* - \sum_{j=1}^m x^*_{i_j}$. Then by (5.9), $\hat{y}^* \in Z^\perp$ and

$$(y^*, \alpha) = (\hat{y}^*, 0) + (v^*, \hat{\beta}) + \sum_{j=1}^m \left(x^*_{i_j}, \hat{\alpha}_{i_j}\right) \in Z^\perp \times \{0\} + \operatorname{epi} \sigma_D + \sum_{j=1}^m \operatorname{epi} \sigma_{C_{i_j} \cap Z}.$$

Thus, (5.4) holds, as $Z^\perp \times \{0\}$ is clearly contained in $\operatorname{epi} \sigma_D$. □

Remark 5.1. If, for (a) of Lemma 5.2, $\dim D \le m-1$, then the following implication is valid:

$$\text{(a)} \wedge \text{(b)} \wedge \text{(c)} \Rightarrow \{D, (\operatorname{span} D) \cap C_i : i \in I\} \text{ satisfies the SECQ.}$$

(This can be seen from (i) of Theorem 5.3 below, but with m replaced by $m-1$.)

THEOREM 5.3. *Let $m \in \mathbb{N}$ and let $\{D, C_i : i \in I\}$ be a CCS-system with the base-set D. We consider the following conditions:*

(a) *D is of finite dimension m.*
(b) *The set-valued mapping $i \mapsto (\operatorname{span} D) \cap C_i$ is lower semicontinuous on I.*
(c) *The system $\{D, C_i : i \in I\}$ satisfies the $(m+1)$-D-interior-point condition.*
(d) *For each $i \in I$, the pair $\{D, C_i\}$ has the property*

$$\operatorname{epi} \sigma_{(\operatorname{span} D) \cap C_i} \subseteq \operatorname{epi} \sigma_D + \operatorname{epi} \sigma_{C_i}$$

(e.g., $\{D, C_i\}$ satisfies the SECQ).
(c*) *The system $\{D, C_i : i \in I\}$ satisfies the m-D-interior-point condition.*
(d*) *For each finite subset J of I with $|J| = \min\{m+1, |I|\}$, the subsystem $\{D, C_j : j \in J\}$ satisfies the SECQ.*

Then the following assertions hold:

(i) *If* (a), (b), (c) *are satisfied, then $\{D, (\operatorname{span} D) \cap C_i : i \in I\}$ satisfies the SECQ.*
(ii) *If* (a), (b), (c), (d) *are satisfied, then $\{D, C_i : i \in I\}$ satisfies the SECQ.*
(iii) *If D is bounded and* (a), (b), (c*), (d*) *are satisfied, then $\{D, C_i : i \in I\}$ satisfies the SECQ.*

Proof. (i) Write $Z := \operatorname{span} D$ as before. For a subset H of $X^* \times \mathbb{R}$, we use $H|_Z \subseteq Z^* \times \mathbb{R}$ to denote the restriction to Z of H defined by

$$H|_Z = \{(x^*|_Z, \beta) : \ (x^*, \beta) \in H\}.$$

Let $(y^*, \alpha) \in \overline{\operatorname{epi}\sigma_D + \sum_{i\in I} \operatorname{epi}\sigma_{C_i}}^{w^*}$. Since Z is finite dimensional, there exists a sequence $\{(y_k^*, \alpha_k)\} \subseteq X^* \times \mathbb{R}$ with

$$(y_k^*, \alpha_k) \in \operatorname{epi}\sigma_D + \sum_{i\in I} \operatorname{epi}\sigma_{C_i} \quad \text{for each } k \in \mathbb{N} \tag{5.10}$$

such that $(y_k^*|_Z, \alpha_k)$ converges to $(y^*|_Z, \alpha)$. By (5.10) we express for each $k \in \mathbb{N}$,

$$(y_k^*, \alpha_k) = (v_k^*, \beta_k) + (u_k^*, \gamma_k), \tag{5.11}$$

where $(v_k^*, \beta_k) \in \operatorname{epi}\sigma_D$ and $(u_k^*, \gamma_k) \in \sum_{i\in I} \operatorname{epi}\sigma_{C_i}$. Since $(\sum_{i\in I} \operatorname{epi}\sigma_{C_i})|_Z$ is a convex cone in the $(m+1)$-dimensional space $Z^* \times \mathbb{R}$, it follows from [32, Theorem 3.15] that, for each k, there exist indices $\{i_1^k, \ldots, i_{m+1}^k\} \subseteq I$ and $\{(x_{i_1^k}^*, \alpha_{i_1^k}), \ldots, (x_{i_{m+1}^k}^*, \alpha_{i_{m+1}^k})\}$ with $(x_{i_j^k}^*, \alpha_{i_j^k}) \in \operatorname{epi}\sigma_{C_{i_j^k}}$ for each $1 \le j \le m+1$ such that

$$(u_k^*|_Z, \gamma_k) = \sum_{j=1}^{m+1} \left(x_{i_j^k}^*\Big|_Z, \alpha_{i_j^k} \right) \quad \text{for each } k \in \mathbb{N}. \tag{5.12}$$

Thus we have

$$(y_k^*|_Z, \alpha_k) = (v_k^*|_Z, \beta_k) + \sum_{j=1}^{m+1} \left(x_{i_j^k}^*\Big|_Z, \alpha_{i_j^k} \right) \quad \text{for each } k \in \mathbb{N}. \tag{5.13}$$

By Lemma 5.2 and thanks to assumptions (a), (b), (c),

$$(y^*, \alpha) \in \operatorname{epi}\sigma_D + \sum_{i\in I} \operatorname{epi}\sigma_{Z\cap C_i}.$$

We have just proved the inclusion

$$\overline{\operatorname{epi}\sigma_D + \sum_{i\in I} \operatorname{epi}\sigma_{C_i}}^{w^*} \subseteq \operatorname{epi}\sigma_D + \sum_{i\in I} \operatorname{epi}\sigma_{Z\cap C_i}. \tag{5.14}$$

Noting $D \cap (\cap_{i\in I}(Z \cap C_i)) = D \cap (\cap_{i\in I} C_i)$, it follows from Proposition 2.4 and (5.14) that

$$\operatorname{epi}\sigma_{D\cap(\cap_{i\in I}(Z\cap C_i))} = \overline{\operatorname{epi}\sigma_D + \sum_{i\in I} \operatorname{epi}\sigma_{C_i}}^{w^*} \subseteq \operatorname{epi}\sigma_D + \sum_{i\in I} \operatorname{epi}\sigma_{Z\cap C_i}. \tag{5.15}$$

Thus $\{D, (\operatorname{span} D) \cap C_i : \ i \in I\}$ satisfies the SECQ by Corollary 2.5. This proves assertion (i).

(ii) Now suppose in addition that (d) is also satisfied. Then (5.15) implies that

$$\operatorname{epi}\sigma_{D\cap\cap_{i\in I}C_i} \subseteq \operatorname{epi}\sigma_D + \sum_{i\in I}(\operatorname{epi}\sigma_D + \operatorname{epi}\sigma_{C_i}) \subseteq \operatorname{epi}\sigma_D + \sum_{i\in I} \operatorname{epi}\sigma_{C_i}.$$

By Corollary 2.5 again, this implies that $\{D, C_i : i \in I\}$ satisfies the SECQ; that is, (ii) holds.

(iii) Now suppose that (a), (b), (c*), (d*) are satisfied. Without loss of generality, we may assume that $|I| > m+1$ since, otherwise, the conclusion follows from assumption (d*). Consider (y^*, α), (y_k^*, α_k), (v_k^*, β_k), (u_k^*, γ_k) satisfying (5.10)–(5.13). Let $k \in \mathbb{N}$ and set $I^k = \{i_1^k, \ldots, i_{m+1}^k\}$. Then for any $z \in D \cap \cap_{j\in I^k} C_j$ $(\subseteq Z)$,

$$\alpha_k = \beta_k + \sum_{j\in I^k} \alpha_{i_j^k} \geq \sigma_D(v_k^*) + \sum_{j=1}^{m+1} \sigma_{C_j}\left(x_{i_j^k}^*\right) \geq \left\langle v_k^* + \sum_{j=1}^{m+1} x_{i_j^k}^*, z \right\rangle = \langle y_k^*, z\rangle,$$

thanks to (5.13). Since $D \cap (\bigcap_{j\in I^k} C_j)$ is compact, there exists $x^k \in D \cap (\bigcap_{j\in I^k} C_j)$ such that

$$\alpha_k \geq \langle y_k^*, x^k\rangle = \sigma_{D\cap(\bigcap_{j\in I^k} C_j)}(y_k^*), \tag{5.16}$$

i.e., $y_k^* \in N_{D\cap(\cap_{j\in I^k} C_j)}(x^k)$. It follows from assumption (d*) and Theorem 3.1 that $y_k^* \in N_D(x^k) + \sum_{j\in I^k} N_{C_j}(x^k)$. Applying [32, Theorem 3.15] to the m-dimensional linear subspace Z, $y_k^*|_Z$ can be expressed in the form

$$y_k^*|_Z = d_k^*|_Z + \sum_{j\in J^k} z_j^*|_Z \tag{5.17}$$

for some $d_k^* \in N_D(x^k)$ and $z_j^* \in N_{C_j}(x^k)$ $(j \in J^k)$, where J^k is a subset of I^k with m elements. Evaluating (5.17) at $x^k \in D \cap (\cap_{j\in I^k} C_j)$, and invoking (2.4) and (5.16), we have

$$\alpha_k \geq \langle y_k^*, x^k\rangle = \sigma_D(d_k^*) + \sum_{j\in J^k} \sigma_{C_j}(z_j^*). \tag{5.18}$$

Define

$$\mu_k = \alpha_k - \sum_{j\in J^k} \sigma_{C_j}(z_j^*).$$

Then $\mu_k \geq \sigma_D(d_k^*)$ by (5.18). Denoting $\sigma_{C_j}(z_j^*)$ by γ_j, this and (5.17) imply that

$$(y_k^*|_Z, \alpha_k) = (d_k^*|_Z, \mu_k) + \sum_{j\in J^k} (z_j^*|_Z, \gamma_j).$$

Note that $(d_k^*, \mu_k) \in \operatorname{epi}\sigma_D$ and $(z_j^*, \gamma_j) \in \operatorname{epi}\sigma_{C_j}$ for each $j \in J^k$. Since $|J^k| = m$ and thanks to assumptions (a), (b), and (c*), Lemma 5.2 asserts that

$$(y^*, \alpha) \in \operatorname{epi}\sigma_D + \sum_{i\in I} \operatorname{epi}\sigma_{(Z\cap C_i)}. \tag{5.19}$$

Let $i \in I$ and let J be any subset of I such that $i \in J$ and $|J| = m+1$. Then, by assumption (d*), one has that

$$\operatorname{epi}\sigma_{(Z\cap C_i)} \subseteq \operatorname{epi}\sigma_{(D\cap(\cap_{j\in J} C_j))} \subseteq \operatorname{epi}\sigma_D + \sum_{j\in J} \operatorname{epi}\sigma_{C_j}. \tag{5.20}$$

Therefore, by (5.19) and (5.20), $(y^*, \alpha) \in \operatorname{epi}\sigma_D + \sum_{i\in I}\operatorname{epi}\sigma_{C_i}$, and thus $\operatorname{epi}\sigma_D + \sum_{i\in I}\operatorname{epi}\sigma_{C_i}$ is weakly* closed in the case when assumptions (a), (b), (c*), and (d*) are satisfied. By Corollary 2.5, this implies that $\{D, C_i : i \in I\}$ satisfies the SECQ. The proof is complete. □

COROLLARY 5.4. *Let $m \in \mathbb{N}$ and let $\{D, C_i : i \in I\}$ be a CCS-system with the base-set D satisfying the following conditions:*

(a) *D is of finite dimension m.*
(b) *The set-valued mapping $i \mapsto (\operatorname{span} D) \cap C_i$ is lower semicontinuous on I.*
(c$^+$) *The system $\{D, C_i : i \in I\}$ satisfies the $(m+1)$-interior-point condition.*

Then $\{D, C_i : i \in I\}$ satisfies the SECQ.

Proof. By Lemma 4.9, (c$^+$) implies conditions (d) and (c) of Theorem 5.3. Thus, Theorem 5.3(ii) is applicable. □

The following corollary, which is a direct consequence of Theorem 5.3(i), is an improvement of Theorem 1.1.

COROLLARY 5.5. *Let $\{D, C_i : i \in I\}$ be a CCS-system with the base-set D. Let $m \in \mathbb{N}$ and let $x_0 \in D \cap C$. Suppose that the following conditions are satisfied:*

(a) *D is of finite dimension m.*
(b) *The set-valued mapping $i \mapsto (\operatorname{span} D) \cap C_i$ is lower semicontinuous on I.*
(c) *The system $\{D, C_i : i \in I\}$ satisfies the $(m+1)$-D-interior-point condition.*
(d̃) *For each $i \in I$, the pair $\{D, C_i\}$ has the property*

$$N_{(\operatorname{span} D)\cap C_i}(x_0) \subseteq N_D(x_0) + N_{C_i}(x_0).$$

Then the system $\{D, C_i : i \in I\}$ has the strong CHIP at x_0.

The following corollary is an important improvement of Theorem 1.2. Our main improvement lies in the fact that we need not require the upper semicontinuity of the set-valued map $i \mapsto (\operatorname{span} D) \cap C_i$ and that (d) can be weakened to require that only the subsystems $\{D, C_j : j \in J\}$ with $|J| = l + 1$ have the strong CHIP.

COROLLARY 5.6. *Let $m \in \mathbb{N}$ and let $\{D, C_i : i \in I\}$ be a CCS-system with the base-set D satisfying the following conditions:*

(a) *D is of finite dimension m.*
(b) *The set-valued mapping $i \mapsto (\operatorname{span} D) \cap C_i$ is lower semicontinuous on I.*
(c*) *The system $\{D, C_i : i \in I\}$ satisfies the m-D-interior-point condition.*
(d) *For each finite subset J of I with $|J| = \min\{m+1, |I|\}$, the subsystem $\{D, C_j : j \in J\}$ has the strong CHIP.*

Then the system $\{D, C_i : i \in I\}$ has the strong CHIP.

Proof. If $|I| < m + 1$, then $\min\{m + 1, |I|\} = |I|$, so the result is trivially true by (d). Thus we may assume that $|I| \geq m + 1$. Recall that $C = \cap_{i\in I} C_i$ and let $x \in D \cap C$. We have to show that the system has the strong CHIP at x. To this end, let $\tilde{D} = D \cap \mathbf{B}(x, r_x)$, where $r_x = \|x\| + 1$. Consider the system $\{\tilde{D}, C_i : i \in I\}$. We claim that the following conditions hold:

(ã) $\tilde{D}$ is of finite dimension and $\dim \tilde{D} = m$.
(b̃) The set-valued mapping $i \mapsto (\operatorname{span} \tilde{D}) \cap C_i$ is lower semicontinuous on I.
(c̃) The system $\{\tilde{D}, C_i : i \in I\}$ satisfies the m-$\tilde{D}$-interior-point condition.
(d̃) For each finite subset J of I with $|J| = m+1$, the subsystem $\{\tilde{D}, C_j : j \in J\}$ satisfies the SECQ.

In fact, by assumption (c*), for each finite subset J of I with $|J| = m$, there exist $\bar{x} \in D$ and $\delta > 0$ such that $\mathbf{B}(\bar{x}, \delta) \cap \operatorname{span} D \subseteq D \cap (\cap_{j\in J} C_j)$. Since $0 \in \operatorname{int} \mathbf{B}(x, r_x)$, there exists $\lambda \in (0, 1)$ such that $\lambda\mathbf{B}(\bar{x}, \delta) \subseteq \mathbf{B}(x, r_x)$. Consequently,

$$\lambda \mathbf{B}(\bar{x},\delta) \bigcap \operatorname{span} D \subseteq D \bigcap \mathbf{B}(x, r_x) \bigcap \left(\bigcap_{j\in J} C_j \right) \subseteq \tilde{D} \bigcap \left(\bigcap_{i\in J} C_i \right). \tag{5.21}$$

This implies that $\operatorname{int} \mathbf{B}(\bar{x},\delta) \cap \operatorname{ri} D \neq \emptyset$; hence

$$\operatorname{span} \tilde{D} = \operatorname{span} D. \tag{5.22}$$

Consequently, condition (c̃) holds by (5.21). Moreover, by (a), (b), and (5.22), it is seen that (ã) and (b̃) hold. As to condition (d̃), let J be any subset of I with $|J| = m + 1$. By (d) the subsystem $\{D, C_j : j \in J\}$ has the strong CHIP. Since $x \in \operatorname{int} \mathbf{B}(x, r_x) \cap (D \cap (\cap_{j\in J} C_j))$, and by applying Lemma 4.12 to the ball with center x, radius r_x, and J in place of A and I, it follows that $\{\tilde{D}, C_j : j \in J\}$ has the strong CHIP and consequently satisfies the SECQ, thanks to Corollary 3.2(i) because $\tilde{D} \cap (\cap_{j\in J} C_j)$ is compact. Thus (d̃) is established. Thus part (iii) of Theorem 5.3 is applicable for concluding that the system $\{\tilde{D}, C_i : i \in I\}$ satisfies the SECQ, which in turn implies that it has the strong CHIP at x. Consequently, the system has the strong CHIP at x by Lemma 4.12 applied to the ball with center x, radius r_x, and J in place of A and I. The proof is complete. □

Acknowledgments. The authors would like to express their sincere thanks to the two referees for many helpful comments and for shorter proofs of Theorem 4.6 and of other minor results.

REFERENCES

[1] J. P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhäuser Boston, Boston, MA, 1990.
[2] A. Auslender and M. Teboulle, *Asymptotic Cones and Functions in Optimization and Variational Inequalities*, Springer-Verlag, New York, 2003.
[3] A. Bakan, F. Deutsch, and W. Li, *Strong CHIP, normality, and linear regularity of convex sets*, Trans. Amer. Math. Soc., 357 (2005), pp. 3831–3863.
[4] H. Bauschke, *Projection Algorithms and Monotone Operators*, Ph.D. thesis, Department of Mathematics, Simon Fraser University, Burnaby, BC, Canada, 1996. Available online at http://oldweb.cecm.sfu.ca/preprints/1996pp.html.
[5] H. H. Bauschke and J. M. Borwein, *On projection algorithms for solving convex feasibility problems*, SIAM Rev., 38 (1996), pp. 367–426.
[6] H. Bauschke, J. Borwein, and W. Li, *Strong conical hull intersection property, bounded linear regularity, Jameson's property (G), and error bounds in convex optimization*, Math. Program. Ser. A, 86 (1999), pp. 135–160.
[7] R. S. Burachik and V. Jeyakumar, *A simple closure condition for the normal cone intersection formula*, Proc. Amer. Math. Soc., 133 (2005), pp. 1741–1748.
[8] F. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley & Sons, New York, 1983.
[9] F. Deutsch, *The role of the strong conical hull intersection property in convex optimization and approximation*, in Approximation Theory IX, Vol. I: Theoretical Aspects, C. Chui and L. Schumaker, eds., Vanderbilt University Press, Nashville, TN, 1998, pp. 105–112.
[10] F. Deutsch, *Best Approximation in Inner Product Spaces*, Springer, New York, 2001.
[11] F. Deutsch, W. Li, and J. Swetits, *Fenchel duality and the strong conical hull intersection property*, J. Optim. Theory Appl., 102 (1997), pp. 681–695.
[12] F. Deutsch, W. Li, and J. Ward, *A dual approach to constrained interpolation from a convex subset of Hilbert space*, J. Approx. Theory, 90 (1997), pp. 385–414.
[13] F. Deutsch, W. Li, and J. D. Ward, *Best approximation from the intersection of a closed convex set and a polyhedron in Hilbert space, weak Slater conditions, and the strong conical hull intersection property*, SIAM J. Optim., 10 (1999), pp. 252–268.
[14] D. Gale and V. Klee, *Continuous convex sets*, Math. Scand., 7 (1959), pp. 370–391.
[15] J. Hiriart-Urruty and C. Lemarechal, *Convex Analysis and Minimization Algorithms* I, Grundlehren Math. Wiss. 305, Springer, New York, 1993.
[16] J. Hiriart-Urruty and C. Lemarechal, *Convex Analysis and Minimization Algorithms* II, Grundlehren Math. Wiss. 306, Springer, New York, 1993.

[17] G. J. O. Jameson, *The duality of pairs of wedges*, Proc. London Math. Soc., 24 (1972), pp. 531–547.
[18] V. Jeyakumar, G. M. Lee, and N. Dinh, *New sequential Lagrange multiplier conditions characterizing optimality without constraint qualification for convex programs*, SIAM J. Optim., 14 (2003), pp. 534–547.
[19] V. Jeyakumar, N. Dinh, and G. M. Lee, *A New Closed Cone Constraint Qualification for Convex Optimization*, Applied Mathematics Research Report AMR 04/6, University of New South Wales, Sydney, Australia.
[20] V. Jeyakumar and H. Mohebi, *A global approach to nonlinearly constrained best approximation*, Numer. Funct. Anal. Optim., 26 (2005), pp. 205–227.
[21] V. Jeyakumar, A. M. Rubinov, B. M. Glover, and Y. Ishizuka, *Inequality systems and global optimization*, J. Math. Anal. Appl., 202 (1996), pp. 900–919.
[22] V. Jeyakumar and A. Zaffaroni, *Asymptotic conditions for weak and proper optimality in infinite dimensional convex vector optimization*, Numer. Funct. Anal. Optim., 17 (1996), pp. 323–343.
[23] A. Lewis and J. S. Pang, *Error bounds for convex inequality systems*, in Generalized Convexity, Generalized Monotonicity: Recent Results, Proceedings of the Fifth Symposium on Generalized Convexity, Luminy, 1996, J.-P. Crouzeix, J.-E. Martinez-Legaz, and M. Volle, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1997, pp. 75–110.
[24] C. Li and X.-Q. Jin, *Nonlinearly constrained best approximation in Hilbert spaces: The strong CHIP and the basic constraint qualification*, SIAM J. Optim., 13 (2002), pp. 228–239.
[25] C. Li and K. F. Ng, *On best approximation by nonconvex sets and perturbation of nonconvex inequality systems in Hilbert spaces*, SIAM J. Optim., 13 (2002), pp. 726–744.
[26] C. Li and K. F. Ng, *Constraint qualification, the strong CHIP, and best approximation with convex constraints in Banach spaces*, SIAM J. Optim., 14 (2003), pp. 584–607.
[27] C. Li and K. F. Ng, *Strong CHIP for infinite system of closed convex sets in normed linear spaces*, SIAM J. Optim., 16 (2005), pp. 311–340.
[28] C. Li and K. F. Ng, *On best restricted range approximation in continuous complex-valued function spaces*, J. Approx. Theory, 136 (2005), pp. 159–181.
[29] J. J. Moreau, *Fonctionnelles convexes*, Séminaire sur les Équations aux dérivées partielles, Collège de France, Paris, 1967.
[30] K. F. Ng and W. H. Yang, *Regularities and their relations to error bounds*, Math. Program. Ser. A, 99 (2004), pp. 521–538.
[31] J. S. Pang, *Error bounds in mathematical programming*, Math. Program. Ser. B, 79 (1997), pp. 299–332.
[32] R. T. Rockafellar and J. B. Wets, *Variational Analysis*, Springer-Verlag, New York, 1998.
[33] I. Singer, *Duality for optimization and best approximation over finite intersection*, Numer. Funct. Anal. Optim., 19 (1998), pp. 903–915.
[34] W. Song and R. Zang, *Bounded linear regularity of convex sets in Banach spaces and its applications*, Math. Program. Ser. A, 106 (2006), pp. 59–79.
[35] C. Zălinescu, *Convex Analysis in General Vector Spaces*, World Scientific, River Edge, NJ, 2002.

SIAM J. OPTIM.
Vol. 18, No. 2, pp. 666–690

ITERATIVE SOLUTION OF AUGMENTED SYSTEMS ARISING IN INTERIOR METHODS*

ANDERS FORSGREN†, PHILIP E. GILL‡, AND JOSHUA D. GRIFFIN§

Abstract. Iterative methods are proposed for certain augmented systems of linear equations that arise in interior methods for general nonlinear optimization. Interior methods define a sequence of *KKT equations* that represent the symmetrized (but indefinite) equations associated with Newton's method for a point satisfying the perturbed optimality conditions. These equations involve both the primal and dual variables and become increasingly ill-conditioned as the optimization proceeds. In this context, an iterative linear solver must not only handle the ill-conditioning but also detect the occurrence of KKT matrices with the wrong matrix inertia. A one-parameter family of equivalent linear equations is formulated that includes the KKT system as a special case. The discussion focuses on a particular system from this family, known as the "doubly augmented system," that is positive definite with respect to both the primal and dual variables. This property means that a standard preconditioned conjugate-gradient method involving both primal and dual variables will either terminate successfully or detect if the KKT matrix has the wrong inertia. Constraint preconditioning is a well-known technique for preconditioning the conjugate-gradient method on augmented systems. A family of constraint preconditioners is proposed that provably eliminates the inherent ill-conditioning in the augmented system. A considerable benefit of combining constraint preconditioning with the doubly augmented system is that the preconditioner need not be applied exactly. Two particular "active-set" constraint preconditioners are formulated that involve only a subset of the rows of the augmented system and thereby may be applied with considerably less work. Finally, some numerical experiments illustrate the numerical performance of the proposed preconditioners and highlight some theoretical properties of the preconditioned matrices.

Key words. large-scale nonlinear programming, nonconvex optimization, interior methods, augmented systems, KKT systems, iterative methods, conjugate-gradient method, constraint preconditioning

AMS subject classifications. 49J20, 49J15, 49M37, 49D37, 65F05, 65K05, 90C30

DOI. 10.1137/060650210

1. Introduction.

This paper concerns the formulation and analysis of preconditioned iterative methods for the solution of augmented systems of the form

$$\begin{pmatrix} H & -A^T \\ A & G \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}, \tag{1.1}$$

with A an $m \times n$ matrix, H symmetric, and G symmetric positive semidefinite. These equations arise in a wide variety of scientific and engineering applications, where they are known by a number of different names, including "augmented systems," "saddle-point systems," "KKT systems," and "equilibrium systems." (The bibliography of the survey by Benzi, Golub, and Liesen [3] contains 513 related articles.) The main focus

*Received by the editors January 17, 2006; accepted for publication (in revised form) March 5, 2007; published electronically August 22, 2007. The research of the second and third authors was supported by National Science Foundation grants DMS-9973276, CCF-0082100, and DMS-0511766.
http://www.siam.org/journals/siopt/18-2/65021.html

†Optimization and Systems Theory, Department of Mathematics, Royal Institute of Technology, SE-100 44 Stockholm, Sweden (andersf@kth.se). The research of this author was supported by the Swedish Research Council (VR).

‡Department of Mathematics, University of California, San Diego, La Jolla, CA 92093-0112 (pgill@ucsd.edu).

§Sandia National Laboratories, Livermore, CA 94551-9217 (jgriffi@sandia.gov). Part of this work was carried out during the Spring of 2003 while this author was visiting KTH with financial support from the Göran Gustafsson Foundation.

of this paper will be on the solution of augmented systems arising in interior methods for general constrained optimization, in which case (1.1) is the system associated with Newton's method for finding values of the primal and dual variables that satisfy the perturbed KKT optimality conditions (see, e.g., Wright [49] and Forsgren, Gill, and Wright [15]). In this context H is the Hessian of the Lagrangian, A is the constraint Jacobian, and G is diagonal.

Many of the benefits associated with the methods discussed in this paper derive from formulating the interior method so that the diagonal G is *positive definite*. We begin by presenting results for G positive definite and consider the treatment of systems with positive semidefinite and singular G in section 5. Throughout, for the case where G is positive definite, we denote G by D and rewrite (1.1) as an equivalent *symmetric* system $Bx = b$, where

$$B = \begin{pmatrix} H & -A^T \\ -A & -D \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} b_1 \\ -b_2 \end{pmatrix}, \tag{1.2}$$

with D positive definite and diagonal. We will refer to this symmetric system as the *KKT system*. (It is possible to symmetrize (1.1) in a number of different ways. The format (1.2) will simplify the linear algebra in later sections.) When D is nonsingular, it is well known that the augmented system is equivalent to the two smaller systems

$$(H + A^T D^{-1} A)x_1 = b_1 + A^T D^{-1} b_2 \quad \text{and} \quad x_2 = D^{-1}(b_2 - Ax_1), \tag{1.3}$$

where the system for x_1 is known as the *condensed system*. It is less well known that another equivalent system is the *doubly augmented system*

$$\begin{pmatrix} H + 2A^T D^{-1} A & A^T \\ A & D \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} b_1 + 2A^T D^{-1} b_2 \\ b_2 \end{pmatrix}, \tag{1.4}$$

which has been proposed for use with direct factorization methods by Forsgren and Gill [16]. In this paper we investigate the properties of preconditioned iterative methods applied to system (1.2) directly or to the equivalent systems (1.3) and (1.4).

If the underlying optimization problem is not convex, the matrix H may be indefinite. The KKT matrix B of (1.2) is said to have *correct inertia* if the matrix $H + A^T D^{-1} A$ is positive definite. This definition is based on the properties of the underlying optimization problem. Broadly speaking, the KKT system has correct inertia if the problem is locally convex (for further details see, e.g., Forsgren and Gill [16], Forsgren [18], and Griffin [32]). If the KKT matrix has correct inertia, then systems (1.2)–(1.4) have a common unique solution (see section 2).

1.1. Properties of the KKT system. The main issues associated with using iterative methods to solve KKT systems are (i) *termination control*, (ii) *inertia control*, and (iii) *inherent ill-conditioning*. The first of these issues is common to other applications where the linear system represents a linearization of some underlying nonlinear system of equations. Issues (ii) and (iii), however, are unique to optimization and will be the principal topics of this paper.

In the context of interior methods, the KKT system (1.2) is solved as part of a two-level iterative scheme. At the outer level, nonlinear equations that define the first-order optimality conditions are parameterized by a small positive quantity μ. The idea is that the solution of the parameterized equations should approach the solution of the optimization problem as $\mu \to 0$. At the inner level, equations (1.2) represent the symmetrized Newton equations associated with finding a zero of the

perturbed optimality conditions for a given value of μ. Although systems (1.2)–(1.4) have identical solutions, an iterative method will generally produce a different sequence of iterates in each case (see section 3 for a discussion of the equivalence of iterative solvers in this context). An iterative method applied to the augmented system (1.2) or the doubly augmented system (1.4) treats x_1 and x_2 as independent variables, which is appropriate in the optimization context because x_1 and x_2 are associated with independent quantities in the perturbed optimality conditions (i.e., the primal and dual variables). In contrast, an iterative solver for the condensed system (1.3) will generate approximations to x_1 only, with the variables x_2 being defined as $x_2 = D^{-1}(b_2 - Ax_1)$. This becomes an important issue when an *approximate* solution is obtained by *truncating* the iterations of the linear solver. During the early outer iterations, it is usually inefficient to solve the KKT system accurately, and it is better to accept an inexact solution that gives a residual norm that is less than some factor of the norm of the right-hand side (see, e.g., Dembo, Eisenstat, and Steihaug [7]). For the condensed system, the residual for the second block of equations will be zero regardless of the accuracy of x_1, which implies that termination must be based on the accuracy of x_1 alone. It is particularly important for the solver to place equal weight on x_1 and x_2 when system (1.2) is being solved in conjunction with a primal-dual trust-region method (see Gertz and Gill [20] and Griffin [32]). The conjugate-gradient version of this method exploits the property that the norms of the (x_1, x_2) iterates increase monotonically (see Steihaug [44]). This property does not hold for (x_1, x_2) iterates generated for the condensed system.

If the KKT matrix does not have the correct inertia, the solution of (1.2) is not useful, and the optimization continues with an alternative technique based on either implicitly or explicitly modifying the matrix H (see, e.g., Toint [45], Steihaug [44], Gould et al. [30], Hager [33], and Griffin [32]). It is therefore important that the iterative solver is able to detect if B does not have correct inertia.

As the perturbation parameter μ is reduced, the KKT systems become *increasingly ill-conditioned*. The precise form of this ill-conditioning depends on the formulation of the interior method, but a common feature is that some diagonal elements of D are big and some are small. (It is almost always possible to formulate an interior method that requires the solution of an *unsymmetric* system that does not exhibit inevitable ill-conditioning as $\mu \to 0$. This unsymmetric system could be solved using an unsymmetric solver such as GMRES or QMR. Unfortunately, this approach is unsuitable for general KKT systems because an unsymmetric solver is unable to determine if the KKT matrix has correct inertia.) In section 3 we consider a preconditioned conjugate-gradient (PCG) method that provably removes the inherent ill-conditioning. In particular, we define a one-parameter family of preconditioners related to the class of so-called *constraint preconditioners* proposed by Keller, Gould, and Wathen [34]. Several authors have used constraint preconditioners in conjunction with the conjugate-gradient method to solve the indefinite KKT system (1.2) with $b_2 = 0$ and $D = 0$ (see, e.g., Lukšan and Vlček [36], Gould, Hribar, and Nocedal [29], Perugia and Simoncini [40], and Bergamaschi, Gondzio, and Zilli [4]). Recently, Dollar [12] and Dollar et al. [11] have proposed constraint preconditioners for system (1.2) with no explicit inertial or diagonal condition on D, but a full row-rank requirement on A and the assumption that $b_2 = 0$.

Methods that require $b_2 = 0$ must perform an initial projection step that effectively shifts the right-hand side to zero. The constraint preconditioner then forces the x_1 iterates to lie in the null space of A. A disadvantage with this approach is that the

constraint preconditioner must be applied exactly if subsequent iterates are to lie in the null space. This limits the ability to perform approximate solves with the preconditioner, as is often required when the matrix A has a PDE-like structure that also must be handled using an iterative solver (see, e.g., Saad [41], Notay [37], Simoncini and Szyld [43], and Elman et al. [14]). In section 3 we consider preconditioners that do not require the assumption that $b_2 = 0$, and hence do not require an accurate solve with the preconditioner.

1.2. A PCG method for the KKT system. The goal of this paper is to formulate iterative methods that not only provide termination control and inertia control, but also eliminate the inevitable ill-conditioning associated with interior methods. All these features are present in an algorithm based on applying a PCG method to the doubly augmented system (1.4). This system is positive definite if the KKT matrix has correct inertia, and gives equal weight to x_1 and x_2 for early terminations. As preconditioner we use the constraint preconditioner

$$P = \begin{pmatrix} M + 2A^TD^{-1}A & A^T \\ A & D \end{pmatrix}, \tag{1.5}$$

where M is an approximation of H such that $M + A^TD^{-1}A$ is positive definite. The equations $Pv = r$ used to apply the preconditioner are solved by exploiting the equivalence of the systems

$$\begin{pmatrix} M + 2A^TD^{-1}A & A^T \\ A & D \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} r_1 \\ r_2 \end{pmatrix}, \tag{1.6a}$$

$$\begin{pmatrix} M & -A^T \\ -A & -D \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} r_1 - 2A^TD^{-1}r_2 \\ -r_2 \end{pmatrix}, \quad \text{and} \tag{1.6b}$$

$$(M + A^TD^{-1}A)v_1 = r_1 - A^TD^{-1}r_2, \;\; v_2 = D^{-1}(r_2 - Av_1) \tag{1.6c}$$

(see section 3). This allows us to compute the solution of (1.6a) by solving either (1.6b) or (1.6c). (The particular choice will depend on the relative efficiency of the methods available to solve the condensed and augmented systems.)

We emphasize that the doubly augmented systems are never formed or factored explicitly. The matrix associated with the doubly augmented equations (1.4) is used only as an operator to define products of the form $v = Bu$. As mentioned above, the equations (1.6a) that apply the preconditioner are solved using either (1.6b) or (1.6c). An important property of the method is that these equations also may be solved using an iterative method. (It is safe to use the augmented or condensed system for the preconditioner equations $Pv = r$ because the inertia of P is guaranteed by the choice of M (see section 3).)

In section 4 we formulate and analyze two variants of the preconditioner (1.5) that exploit the asymptotic behavior of the elements of D. The use of these so-called *active-set preconditioners* may require significantly less work when the underlying optimization problem has more constraints than variables. In section 5, we consider the case where G is positive semidefinite and singular. Finally, in section 6, we present some numerical examples illustrating the properties of the proposed preconditioners.

1.3. Notation and assumptions. Unless explicitly indicated otherwise, $\|\cdot\|$ denotes the vector two-norm or its subordinate matrix norm. The inertia of a real symmetric matrix A, denoted by $\mathrm{In}(A)$, is the integer triple (a_+, a_-, a_0) giving the number of positive, negative, and zero eigenvalues of A. The spectrum of a (possibly unsymmetric) matrix A is denoted by $\mathrm{eig}(A)$. As the analysis concerns matrices

with only real eigenvalues, $\mathrm{eig}(A)$ is regarded as an ordered set, with the least (i.e., "leftmost") eigenvalue, denoted by $\mathrm{eig}_{\min}(A)$, appearing first. The quantity $\sigma_k(A)$ denotes the kth largest singular value of A. Given a positive-definite A, the unique positive-definite X such that $X^2 = A$ is denoted by $A^{1/2}$. Given vectors x_1 and x_2, the column vector consisting of the elements of x_1 augmented by the elements of x_2 is denoted by (x_1, x_2).

When μ is a positive scalar such that $\mu \to 0$, the notation $p = O(\mu)$ means that there exists a constant K such that $|p| \leq K\mu$ for all μ sufficiently small. For a *positive* p, $p = \Omega(1/\mu)$ implies that there exists a constant K such that $1/p \leq K\mu$ for all μ sufficiently small. In particular, $p = O(1)$ means that $|p|$ is bounded, and, for a positive p, $p = \Omega(1)$ means that p is bounded away from zero. For a positive p, the notation $p = \Theta(1)$ is used for the case where both $p = O(1)$ and $p = \Omega(1)$, so that p remains bounded and is bounded away from zero as $\mu \to 0$.

As discussed in section 1.1, we are concerned with solving a sequence of systems of the form (1.2), where the matrices A, H, and D depend implicitly on μ. In particular, A and H are first and second derivatives evaluated at a point depending on μ, and D is an explicit function of μ. The notation defined above allows us to characterize the properties of H, A, and D in terms of their behavior as $\mu \to 0$. Throughout the analysis, it is assumed that the following properties hold:

- (A_1) $\|H\|$ and $\|A\|$ are both $O(1)$.
- (A_2) The row indices of A may be partitioned into disjoint subsets $\mathcal{S}$, $\mathcal{M}$, and $\mathcal{B}$ such that $d_{ii} = O(\mu)$ for $i \in \mathcal{S}$, $d_{ii} = \Theta(1)$ for $i \in \mathcal{M}$, and $d_{ii} = \Omega(1/\mu)$ for $i \in \mathcal{B}$.
- (A_3) If $A_{\mathcal{S}}$ is the matrix of rows of A with indices in $\mathcal{S}$ and $r = \mathrm{rank}(A_{\mathcal{S}})$, then r remains constant as $\mu \to 0$ and $\sigma_r(A_{\mathcal{S}}) = \Theta(1)$.

The second assumption reflects the fact that for μ sufficiently small, some diagonal elements of D are "small," some are "medium," and some are "big."

It is often the case in practice that the equations and variables corresponding to unit rows of A are eliminated directly from the KKT system. This elimination creates no additional nonzero elements and provides a smaller "partially condensed" system with an $\Omega(1/\mu)$ diagonal term added to H. It will be shown that preconditioners for both the full and partially condensed KKT systems depend on the eigenvalues of the same matrix (see Lemmas 3.4 and 3.5). It follows that our analysis also applies to preconditioners defined for the partially condensed system.

2. A parameterized system of linear equations. In this section, it is shown how the indefinite KKT system (1.2) may be embedded in a family of equivalent linear systems, parameterized by a scalar ν. This parameterization facilitates the simultaneous analysis of the three systems (1.2)–(1.4).

DEFINITION 2.1 (the parameterized system). *Let ν denote a scalar. Associated with the KKT equations $Bx = b$ of* (1.2)*, we define the parameterized equations $B(\nu)x = b(\nu)$, with*

$$B(\nu) = \begin{pmatrix} H + (1+\nu)A^T D^{-1} A & \nu A^T \\ \nu A & \nu D \end{pmatrix} \quad \text{and} \quad b(\nu) = \begin{pmatrix} b_1 + (1+\nu)A^T D^{-1} b_2 \\ \nu b_2 \end{pmatrix},$$

where H is symmetric and D is positive definite and diagonal.

The following proposition states the equivalence of the KKT system (1.2) and the parameterized system of Definition 2.1.

PROPOSITION 2.2 (equivalence of the parameterized systems). *Let ν denote a scalar parameter. If $\nu \neq 0$, then the system $Bx = b$ of* (1.2) *and the system $B(\nu)x =$*

$b(\nu)$ of Definition 2.1 *are equivalent, i.e.,* (1.2) *has a solution* (x_1, x_2) *if and only if* (x_1, x_2) *is a solution to* $B(\nu)x = b(\nu)$. *If* $\nu = 0$, *then* (1.2) *has a solution* (x_1, x_2) *if and only if* x_1 *is a solution of* $B(0)x = b(0)$ *and* $x_2 = D^{-1}(b_2 - Ax_1)$.

We are particularly interested in the parameterized system $B(\nu)x = b(\nu)$ for the values $\nu = -1$, $\nu = 0$, and $\nu = 1$. If $\nu = -1$, we obtain the symmetric KKT system (1.2). If $\nu = 0$, we obtain the condensed system

$$(H + A^T D^{-1} A)x_1 = b_1 + A^T D^{-1} b_2 \quad \text{with } x_2 \text{ arbitrary.} \tag{2.1}$$

In this case, x_2 does not appear in the augmented system and must be computed as $x_2 = D^{-1}(b_2 - Ax_1)$. Finally, if $\nu = 1$, we obtain the doubly augmented system

$$\begin{pmatrix} H + 2A^T D^{-1} A & A^T \\ A & D \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} b_1 + 2A^T D^{-1} b_2 \\ b_2 \end{pmatrix}. \tag{2.2}$$

The next result follows from Lemma A.1 of the appendix and gives the inertia of $B(\nu)$ as a function of ν.

PROPOSITION 2.3 (inertia of the parameterized system). *For the matrix* $B(\nu)$ *of Definition* 2.1*, it holds that*

(i) $\text{In}\big(B(\nu)\big) = \text{In}(H + A^T D^{-1} A) + (m, 0, 0)$ *if* $\nu > 0$;
(ii) $\text{In}\big(B(\nu)\big) = \text{In}(H + A^T D^{-1} A) + (0, m, 0)$ *if* $\nu < 0$; *and*
(iii) $\text{In}\big(B(\nu)\big) = \text{In}(H + A^T D^{-1} A) + (0, 0, m)$ *if* $\nu = 0$.

This proposition implies that the inertia of $B(\nu)$ may be determined from the inertia of $H + A^T D^{-1} A$ and the sign of ν. In particular, the result gives the inertia of the three alternate systems in the situation where the inertia of B is correct and the solver can be allowed to continue solving the system. If the inertia of B is correct, then (i) the doubly augmented system is positive definite; (ii) the KKT system has n positive eigenvalues and m negative eigenvalues; and (iii) the condensed system is positive definite (when regarded as a system involving only x_1).

Proposition 2.3 implies that it is not worth applying a conjugate-gradient method to the general indefinite KKT system (1.2) because this method is unable to estimate the *number* of negative eigenvalues of an indefinite matrix. In contrast, the conjugate-gradient method is appropriate for both the doubly augmented system and the condensed system because indefiniteness is immediately indicated by the occurrence of a negative value of $p_j^T C p_j$, where p_i is a conjugate direction and C is either the doubly augmented matrix or the matrix for the condensed system. In other words, the occurrence of a negative value of $p_j^T C p_j$ indicates that the inertia of the system is incorrect and the search for a solution of (1.2) should be abandoned.

3. Constraint preconditioning for the linear equations. The rate of convergence of the conjugate-gradient method may be accelerated by choosing an appropriate symmetric positive-definite preconditioner of the form $P = R^T R$, and applying the conjugate-gradient method to the preconditioned system $R^{-T} B R^{-1} R x = R^{-T} b$. As is well known, the computations may be arranged so that the preconditioner is applied by solving systems of the form $Pv = r$. It is the eigenvalues of the preconditioned matrix $R^{-T} B R^{-1}$ that determine the rate of convergence. As $\text{eig}(R^{-T} B R^{-1}) = \text{eig}(R^{-1} R^{-T} B R^{-1} R) = \text{eig}(P^{-1} B)$, the analysis may be written in terms of $P^{-1} B$ without regard to R. However, it must be emphasized that P must be symmetric positive definite for the standard PCG method to be well defined.

Several authors have suggested constraint preconditioners for (1.2) and (1.3), in which H is replaced by a "simpler" approximation matrix M such that $M + A^T D^{-1} A$ is

positive definite (see, e.g., Keller, Gould, and Wathen [34] and Bergamaschi, Gondzio, and Zilli [4]).

Under certain circumstances, the PCG method may be applied to all three systems and will give identical results in exact arithmetic.

PROPOSITION 3.1. *Assume that $H + A^TD^{-1}A$ is positive definite. Consider the PCG method applied to the KKT system* (1.2)*, the condensed system* (1.3)*, and the doubly augmented system* (1.4) *with preconditioners*

$$\text{(3.1)} \qquad \begin{pmatrix} M & -A^T \\ -A & -D \end{pmatrix}, \quad M + A^TD^{-1}A, \quad \textit{and} \quad \begin{pmatrix} M + 2A^TD^{-1}A & A^T \\ A & D \end{pmatrix},$$

respectively. If $b_2 = 0$, then the PCG method generates the same sequence of iterates for all three systems (where an x_2 iterate for the condensed system is defined as the product of $-D^{-1}A$ and the x_1 iterate).

The first preconditioner of (3.1) is not positive definite, which implies that it does not fit within the conventional PCG framework. However, Proposition 3.1 implies that the PCG method may be applied safely to the KKT system (1.2) in the special situation where $b_2 = 0$. The result is a *projected* PCG method that can be shown to be formally equivalent to the standard method applied to the condensed system; see, e.g., Lukšan and Vlček [36] and Gould, Hribar, and Nocedal [29].

The condition $b_2 = 0$ may be achieved by choosing a special initial point y. In particular, consider the point (y_1, y_2) such that $Ay_1 + Dy_2 = b_2$, and the appropriate preconditioner is used for each system. Let x denote the generic vector of unknowns (the dimension of x will depend on which of the three systems is to be solved). We may, for example, solve $Py = b$, where P is one of the three appropriate preconditioners, or we may set $y_1 = 0$, $y_2 = D^{-1}b_2$. Then use PCG with preconditioner P to solve

$$B\widehat{x} = b - By \quad \text{and set} \quad x = y + \widehat{x}.$$

In general, if $b_2 \neq 0$, it is not safe to apply the PCG method to the indefinite system (1.2). Moreover, the PCG method will usually generate different iterates for the condensed system (1.3) and the doubly augmented system (1.4).

Finally, we note that the condensed system (1.3) and doubly augmented system (1.4) may be viewed as being preconditioned versions of each other, as defined in the following result.

PROPOSITION 3.2. *Consider the PCG method applied to a generic symmetric system $Ax = b$ with symmetric positive-definite preconditioner P and initial iterate $x_0 = 0$. Let L be a nonsingular matrix with the same dimension as A. Then, if the PCG method is applied to $LAL^T\widehat{x} = Lb$ with preconditioner LPL^T and initial iterate $\widehat{x}_0 = 0$, the PCG iterates are related by the transformation $x = L^T\widehat{x}$.*

If we consider the decomposition

$$\begin{pmatrix} M + 2A^TD^{-1}A & A^T \\ A & D \end{pmatrix} = \begin{pmatrix} I & A^TD^{-1} \\ & I \end{pmatrix} \begin{pmatrix} M + A^TD^{-1}A & \\ & D \end{pmatrix} \begin{pmatrix} I & \\ D^{-1}A & I \end{pmatrix},$$

then Proposition 3.2 implies that the doubly augmented system may be viewed as a particular preconditioned version of the condensed system augmented by the diagonal D for the x_2 variables (or vice versa). This is a further illustration that the proposed approach gives equal weight to x_1 and x_2. We prefer to do the analysis in terms of the doubly augmented system because it provides the parameterization based on the scalar parameter ν.

3.1. Properties of the constraint preconditioners. We now embed the preconditioners of (3.1) within a family of preconditioners, parameterized by the scalar ν. This parameterization is analogous to the parameterization of the matrices of Proposition 2.2. The parameterization allows a unified analysis of the three preconditioners given in (3.1).

DEFINITION 3.3 (a parameterized preconditioner). *Associated with the matrix $B(\nu)$ of Definition* 2.1*, we define the preconditioner $P(\nu)$ as*

$$P(\nu) = \begin{pmatrix} M + (1+\nu)A^T D^{-1} A & \nu A^T \\ \nu A & \nu D \end{pmatrix},$$

where M is a symmetric approximation to H such that

(P_1) $\|M\| = O(1)$*;*

(P_2) $M + A^T D^{-1} A$ *is positive definite;*

(P_3) $\mathrm{eig}_{\min}(M + A^T D^{-1} A) = \Omega(1)$.

Given suitable A and D, a matrix M satisfying the conditions of Definition 3.3 may be found, for example, by using a suitable factorization when solving with $P(-1)$ (see Forsgren and Murray [17], Forsgren and Gill [16], and Forsgren [18]).

Proposition 2.3 gives $\mathrm{In}(P(\nu)) = \mathrm{In}(M + A^T D^{-1} A) + \mathrm{In}(\nu D)$. It follows that $P(\nu)$ is nonsingular for $\nu \neq 0$, positive definite for $\nu > 0$, and $\mathrm{In}(P(\nu)) = (n, m, 0)$ for $\nu < 0$. It is straightforward to show that for all nonzero ν, the eigenvalues of $P(\nu)^{-1} B(\nu)$ are real and independent of ν. The first lemma reveals the structure of $P(\nu)^{-1} B(\nu)$.

LEMMA 3.4 (structure of the parameterized preconditioner). *Let $B(\nu)$ and $P(\nu)$ be defined as in Definitions* 2.1 *and* 3.3*, respectively. Then, for $\nu \neq 0$, it holds that*

$$P(\nu)^{-1} B(\nu) = \begin{pmatrix} S & \\ T & I \end{pmatrix} = \begin{pmatrix} I & \\ -D^{-1}A & I \end{pmatrix} \begin{pmatrix} S & \\ & I \end{pmatrix} \begin{pmatrix} I & \\ D^{-1}A & I \end{pmatrix}, \tag{3.2}$$

where S and T are given by

$$S = (M + A^T D^{-1} A)^{-1} (H + A^T D^{-1} A), \tag{3.3a}$$

$$T = D^{-1} A (M + A^T D^{-1} A)^{-1} (M - H). \tag{3.3b}$$

In addition, the spectrum of $P(\nu)^{-1} B(\nu)$ is independent of ν and consists of m unit eigenvalues and the n eigenvalues of $(M + A^T D^{-1} A)^{-1} (H + A^T D^{-1} A)$.

Proof. The expressions for $P(\nu)^{-1} B(\nu)$ follow from the decomposition given in Lemma A.1 of the appendix. The similarity transform (3.2) implies that the spectrum of $P(\nu)^{-1} B(\nu)$ consists of the n eigenvalues of $(M + A^T D^{-1} A)^{-1} (H + A^T D^{-1} A)$ together with m unit eigenvalues. ∎

Next we relate the $O(\mu)$ diagonal elements in D to eigenvalues of size $1 + O(\mu^{1/2})$ in the $(1,1)$ block S of $P(\nu)^{-1} B(\nu)$ from (3.2).

LEMMA 3.5 (eigenvalues of the parameterized preconditioner). *Let M satisfy assumptions* (P_1)–(P_3) *of Definition* 3.3*. Let $A_{\mathcal{S}}$ denote the submatrix of rows of A associated with diagonal elements of D that are $O(\mu)$. Then the eigenvalues of*

$$(M + A^T D^{-1} A)^{-1} (H + A^T D^{-1} A)$$

are all $O(1)$ with at least $\mathrm{rank}(A_{\mathcal{S}})$ *being* $1 + O(\mu^{1/2})$.

Proof. First we show that $(M + A^T D^{-1} A)^{-1}$ has at least $\mathrm{rank}(A_{\mathcal{S}})$ eigenvalues that are $O(\mu)$. Let $m_1 = \mathrm{rank}(A_{\mathcal{S}})$. Without loss of generality it may be assumed that the rows of A are ordered so that the m_1 row indices in $\mathcal{S}$ corresponding to linearly

independent rows of $A_{\mathcal{S}}$ appear first. This implies that A and D may be partitioned as

$$A = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} \quad \text{and} \quad D = \begin{pmatrix} D_1 & \\ & D_2 \end{pmatrix},$$

where A_1 is $m_1 \times n$, and D_1 is $m_1 \times m_1$ with all eigenvalues $O(\mu)$. Then

$$A^T D^{-1} A = A_1^T D_1^{-1} A_1 + A_2^T D_2^{-1} A_2.$$

Consider the singular-value decomposition $A_1 = U \Sigma V^T$, where U and V are orthonormal matrices of dimension $m_1 \times m_1$ and $n \times m_1$, respectively, and Σ is an $m_1 \times m_1$ diagonal matrix. Let $v = Vp$, where p is an arbitrary m_1-vector of unit length. Then

$$v^T A_1^T D_1^{-1} A_1 v = p^T V^T A_1^T D_1^{-1} A_1 V p = p^T \Sigma U^T D_1^{-1} U \Sigma p \geq \sigma_{m_1}^2 \operatorname{eig}_{\min}(D_1^{-1}). \tag{3.4}$$

Assumption (A$_3$) implies that $\sigma_{m_1} = \Theta(1)$. In addition, all eigenvalues of D_1 are $O(\mu)$, and so (3.4) implies that $v^T A_1^T D_1^{-1} A_1 v = \Omega(1/\mu)$. It follows that

$$v^T (M + A^T D^{-1} A) v = v^T M v + v^T A_1^T D_1^{-1} A_1 v + v^T A_2^T D_2^{-1} A_2 v = \Omega(1/\mu), \tag{3.5}$$

since $v^T M v = O(1)$ and $v^T A_2^T D_2^{-1} A_2 v \geq 0$. As p is an arbitrary unit vector such that $v = Vp$, we conclude from (3.5) that there exists an m_1-dimensional subspace of vectors v such that $v^T (M + A^T D^{-1} A) v$ is $\Omega(1/\mu)$. The Courant–Fischer min-max theorem implies that $M + A^T D^{-1} A$ has at least m_1 eigenvalues that are $\Omega(1/\mu)$; see, e.g., [28, Theorem 8.1.2, p. 394]. It follows that there exists an $n \times m_1$ orthonormal matrix Y and $m_1 \times m_1$ diagonal Λ with $\operatorname{eig}_{\min}(\Lambda) = \Omega(1/\mu)$ such that $(M + A^T D^{-1} A) Y = Y \Lambda$. Since $M + A^T D^{-1} A$ is positive definite, it must hold that $(M + A^T D^{-1} A)^{-1} Y = Y \Lambda^{-1}$ and $(M + A^T D^{-1} A)^{-1/2} Y = Y \Lambda^{-1/2}$ with

$$\|(Y^T (M + A^T D^{-1} A)^{-1} Y)\| = O(\mu), \tag{3.6a}$$

$$\|(M + A^T D^{-1} A)^{-1/2} Y\| = O(\mu^{1/2}). \tag{3.6b}$$

Let $E = (M + A^T D^{-1} A)^{-1/2} (H - M) (M + A^T D^{-1} A)^{-1/2}$ and let Z be an $n \times (n - m_1)$ orthonormal matrix such that the columns of Z form a basis for the null space of Y^T. Then $Q = (Y \ Z)$ is orthonormal and E has the same eigenvalues as the matrix

$$Q^T E Q = \begin{pmatrix} Y^T E Y & Y^T E Z \\ Z^T E Y & Z^T E Z \end{pmatrix},$$

where

$$Y^T E Y = Y^T (M + A^T D^{-1} A)^{-1/2} (H - M) (M + A^T D^{-1} A)^{-1/2} Y, \tag{3.7a}$$

$$Y^T E Z = Y^T (M + A^T D^{-1} A)^{-1/2} (H - M) (M + A^T D^{-1} A)^{-1/2} Z, \tag{3.7b}$$

$$Z^T E Z = Z^T (M + A^T D^{-1} A)^{-1/2} (H - M) (M + A^T D^{-1} A)^{-1/2} Z. \tag{3.7c}$$

Then Definition 3.3 and the order estimates (3.6) imply that $\|Y^T E Y\| = O(\mu)$, $\|Z^T E Y\| = O(\mu^{1/2})$, and $\|Z^T E Z\| = O(1)$. Hence, since $\|Z^T E Y\| = O(\mu^{1/2})$, the eigenvalues of E differ by $O(\mu^{1/2})$ from the eigenvalues of $Y^T E Y$ together with the

eigenvalues of Z^TEZ. But, since $\|Y^TEY\| = O(\mu)$, we conclude that E has m_1 eigenvalues that are $O(\mu^{1/2})$. By similarity,

$$\operatorname{eig}\big((M + A^TD^{-1}A)^{-1}(H - M)\big) = \operatorname{eig}(E),$$

and it must hold that $(M + A^TD^{-1}A)^{-1}(H - M)$ has at least m_1 eigenvalues that are $O(\mu^{1/2})$. The required results now follow from the identity

$$(M + A^TD^{-1}A)^{-1}(H + A^TD^{-1}A) = I + (M + A^TD^{-1}A)^{-1}(H - M),$$

completing the proof. ∎

If $M - H$ is known to be a definite matrix, then the $O(\mu^{1/2})$ bound of Lemma 3.5 may be sharpened to be $O(\mu)$. In this case, the $O(\mu)$ curvature of the product $(M + A^TD^{-1}A)^{-1/2}(H-M)(M+A^TD^{-1}A)^{-1/2}$ over a $\operatorname{rank}(A_{\mathcal{S}})$-dimensional space implied by (3.7a) is sufficient to guarantee $\operatorname{rank}(A_{\mathcal{S}})$ eigenvalues $1 + O(\mu)$.

A combination of Lemmas 3.4 and 3.5 gives the following result on the eigenvalues of $P(\nu)^{-1}B(\nu)$.

THEOREM 3.6 (eigenvalues of the preconditioned matrix). *Let $\nu \neq 0$, and let $B(\nu)$ and $P(\nu)$ be defined as in Definitions* 2.1 *and* 3.3*, respectively. Let $A_{\mathcal{S}}$ denote the submatrix of rows of A associated with diagonal elements of D that are $O(\mu)$. The preconditioned matrix $P(\nu)^{-1}B(\nu)$ has the following properties*:

(a) *The spectrum of $P(\nu)^{-1}B(\nu)$ is independent of ν and consists of m unit eigenvalues and the n eigenvalues of $(M + A^TD^{-1}A)^{-1}(H + A^TD^{-1}A)$.*

(b) *Every eigenvalue of $P(\nu)^{-1}B(\nu)$ is of order $O(1)$. Moreover, $P(\nu)^{-1}B(\nu)$ has at least $m + \operatorname{rank}(A_{\mathcal{S}})$ eigenvalues $1 + O(\mu^{1/2})$, of which at least m are exactly one.*

This result implies that if $m_{\mathcal{S}}$ denotes the number of eigenvalues of D that are $O(\mu)$ and the corresponding $m_{\mathcal{S}} \times n$ submatrix $A_{\mathcal{S}}$ has full row rank, then the PCG method can be expected to give a solution that is $O(\mu^{1/2})$ accurate in at most $n - m_{\mathcal{S}}$ iterations.

4. Active-set preconditioning. An advantage of interior methods is that all inequality constraints are treated in the same way—i.e., the solution path does not depend on an explicit prediction of which constraints are active at the solution. However, this advantage also can be a weakness because all constraint gradients are included in the linear system, even those having little or no influence on the solution. For example, if an interior method is applied to a problem with 100 variables and 100,000 inequality constraints, then a KKT system with 100,100 rows and columns must be solved at each iteration. However, if only 50 (say) of the inequalities are active at the solution, an active-set method would need to solve a KKT system of order 150. In the context of an interior method, the partition of constraints into "active" and "inactive" is determined by the magnitude of the diagonals of D in the KKT system (1.2). Broadly speaking, the active set at the solution is estimated by the indices of the "small" diagonals, and the inactive set is estimated by the indices of the "big" diagonals.

In this section we formulate and analyze two *active-set* preconditioners based on discarding rows of A that correspond to the big diagonals of D. The preconditioners may be applied with a cost comparable to that of solving the KKT system in an active-set method. In addition, the preconditioners allow considerable flexibility in how the diagonals are partitioned into large and small elements—the partition affects

only the rate of convergence of the iterative solver, not the rate of convergence of the interior method. Similar preconditioners have been proposed by Gertz and Griffin [21] in the context of support vector machine classifiers for large data sets. Preconditioners for the solution of linear programs in standard form have been considered by Gill et al. [25] and Oliveira and Sorensen [38]. An active-set preconditioner for general nonlinear optimization has been proposed by Lukšan, Matonoha, and Vlček [35].

4.1. Two active-set preconditioners. Let $m_{\mathcal{S}}$, $m_{\mathcal{M}}$, and $m_{\mathcal{B}}$ denote the number of row indices in the sets $\mathcal{S}$, $\mathcal{M}$, and $\mathcal{B}$ of "small," "medium," and "big" elements of D (see section 1.3). These sets are disjoint, and together they contain all the row indices of A, so that $m_{\mathcal{S}} + m_{\mathcal{M}} + m_{\mathcal{B}} = m$. If strict complementarity holds for the underlying optimization problem, then $m_{\mathcal{M}}$ is zero for all μ sufficiently small (see, e.g., Forsgren, Gill, and Wright [15, p. 531]). The following analysis, does not assume strict complementarity and so $m_{\mathcal{M}}$ may be nonzero. However, it must be emphasized that in this situation, the assumption regarding the order of the small elements of D is a simplification of the real situation. Our assumption that $d_{ii} = O(\mu)$ for $i \in \mathcal{S}$ is sufficient to capture the behavior as μ converges to zero. For a detailed discussion regarding interior methods on degenerate problems, see, e.g., Wright and Orban [48].

In order to simplify the notation, the indices corresponding to the small and medium diagonals are combined into one set $\mathcal{C}$, i.e., $\mathcal{C} = \mathcal{S} \cup \mathcal{M}$. This set is the complement of $\mathcal{B}$, i.e., $\mathcal{C} \cap \mathcal{B} = \emptyset$ and $\mathcal{C} \cup \mathcal{B} = \{1, \ldots, m\}$. This simplification is possible because of our focus on preconditioners based on discarding information associated with the indices in $\mathcal{B}$. Given the partition induced by $\mathcal{B}$ and $\mathcal{C}$, the matrix $P(\nu)$ may be partitioned as

$$P(\nu) = \begin{pmatrix} M + (1+\nu)A^T D^{-1} A & \nu A_{\mathcal{C}}^T & \nu A_{\mathcal{B}}^T \\ \nu A_{\mathcal{C}} & \nu D_{\mathcal{C}} & \\ \nu A_{\mathcal{B}} & & \nu D_{\mathcal{B}} \end{pmatrix}. \tag{4.1}$$

By eliminating the $\nu D_{\mathcal{B}}$ block from $P(\nu)$, we may factor $P(\nu)$ as $P(\nu) = R_{\mathcal{P}} P_{\mathcal{P}}(\nu) R_{\mathcal{P}}^T$, with

$$R_{\mathcal{P}} = \begin{pmatrix} I & & A_{\mathcal{B}}^T D_{\mathcal{B}}^{-1} \\ & I & \\ & & I \end{pmatrix}, \tag{4.2a}$$

$$P_{\mathcal{P}}(\nu) = \begin{pmatrix} M + A_{\mathcal{B}}^T D_{\mathcal{B}}^{-1} A_{\mathcal{B}} + (1+\nu) A_{\mathcal{C}}^T D_{\mathcal{C}}^{-1} A_{\mathcal{C}} & \nu A_{\mathcal{C}}^T & \\ \nu A_{\mathcal{C}} & \nu D_{\mathcal{C}} & \\ & & \nu D_{\mathcal{B}} \end{pmatrix}. \tag{4.2b}$$

Here, the subscript "$\mathcal{P}$" identifies matrices that depend on the partition induced by $\mathcal{B}$ and $\mathcal{C}$ (note that $P(\nu)$ itself is independent of the partition).

The nontrivial step associated with applying the preconditioner in the factored form $P(\nu) = R_{\mathcal{P}} P_{\mathcal{P}}(\nu) R_{\mathcal{P}}^T$ requires a solve with the leading principal submatrix of (4.2b):

$$\begin{pmatrix} M + A_{\mathcal{B}}^T D_{\mathcal{B}}^{-1} A_{\mathcal{B}} + (1+\nu) A_{\mathcal{C}}^T D_{\mathcal{C}}^{-1} A_{\mathcal{C}} & \nu A_{\mathcal{C}}^T \\ \nu A_{\mathcal{C}} & \nu D_{\mathcal{C}} \end{pmatrix}. \tag{4.3}$$

This matrix, formed by eliminating the block $\nu D_{\mathcal{B}}$ from $P(\nu)$, has smaller dimension, $(n + m - m_{\mathcal{B}}) \times (n + m - m_{\mathcal{B}})$, compared to $(n + m) \times (n + m)$ for $P(\nu)$. Lukšan, Matonoha, and Vlček [35] propose an active-set preconditioner based on forming an

incomplete factorization of the $(1,1)$ block. This avoids unnecessary fill-in from the term $A_{\mathcal{B}}^T D_{\mathcal{B}}^{-1} A_{\mathcal{B}}$. We propose an alternative strategy based on the observation that since $\|D_{\mathcal{B}}^{-1}\| = O(\mu)$, then $\|A_{\mathcal{B}}^T D_{\mathcal{B}}^{-1} A_{\mathcal{B}}\| = O(\mu)$. In particular, the term $A_{\mathcal{B}}^T D_{\mathcal{B}}^{-1} A_{\mathcal{B}}$ may be omitted from the $(1,1)$ block of $P_{\mathcal{P}}(\nu)$ without significantly changing the preconditioner. This implies that (4.3) is replaced by $P_{\mathcal{C}}(\nu)$, where

$$P_{\mathcal{C}}(\nu) = \begin{pmatrix} M + (1+\nu) A_{\mathcal{C}}^T D_{\mathcal{C}}^{-1} A_{\mathcal{C}} & \nu A_{\mathcal{C}}^T \\ \nu A_{\mathcal{C}} & \nu D_{\mathcal{C}} \end{pmatrix}. \tag{4.4}$$

The subscript "$\mathcal{C}$" indicates that $P_{\mathcal{C}}(\nu)$ depends only on the indices in $\mathcal{C}$. In active-set constraint preconditioning, $P_{\mathcal{C}}(\nu)$ plays the role of $P(\nu)$ in the analysis of the standard case in section 3.1. Analogous to the assumptions on M, A, and D in Definition 3.3, we require that

- (P'_1) $\|M\| = O(1)$;
- (P'_2) $M + A_{\mathcal{C}}^T D_{\mathcal{C}}^{-1} A_{\mathcal{C}}$ is positive definite; and
- (P'_3) $\mathrm{eig}_{\min}(M + A_{\mathcal{C}}^T D_{\mathcal{C}}^{-1} A_{\mathcal{C}}) = \Omega(1)$.

When $P_{\mathcal{C}}(\nu)$ replaces the leading principal submatrix in (4.2b) the product of the factors becomes

$$P_{\mathcal{P}}^1(\nu) = \begin{pmatrix} M + \nu A_{\mathcal{B}}^T D_{\mathcal{B}}^{-1} A_{\mathcal{B}} + (1+\nu) A_{\mathcal{C}}^T D_{\mathcal{C}}^{-1} A_{\mathcal{C}} & \nu A_{\mathcal{C}}^T & \nu A_{\mathcal{B}}^T \\ \nu A_{\mathcal{C}} & \nu D_{\mathcal{C}} & \\ \nu A_{\mathcal{B}} & & \nu D_{\mathcal{B}} \end{pmatrix}, \tag{4.5}$$

which alternatively may be viewed as the preconditioner obtained by subtracting the term $A_{\mathcal{B}}^T D_{\mathcal{B}}^{-1} A_{\mathcal{B}}$ from the $(1,1)$ block of $P(\nu)$.

The preconditioner (4.5) has the factorization $P_{\mathcal{P}}^1(\nu) = R_{\mathcal{P}} P_{\mathcal{P}}^2(\nu) R_{\mathcal{P}}^T$, where $R_{\mathcal{P}}$ is the upper-triangular matrix (4.2a) and $P_{\mathcal{P}}^2(\nu)$ is given by

$$P_{\mathcal{P}}^2(\nu) = \begin{pmatrix} M + (1+\nu) A_{\mathcal{C}}^T D_{\mathcal{C}}^{-1} A_{\mathcal{C}} & \nu A_{\mathcal{C}}^T & \\ \nu A_{\mathcal{C}} & \nu D_{\mathcal{C}} & \\ & & \nu D_{\mathcal{B}} \end{pmatrix}. \tag{4.6}$$

The matrix $P_{\mathcal{P}}^2(\nu)$ is yet another active-set preconditioner, which may be derived differently by replacing the leading principal submatrix of $P_{\mathcal{P}}(\nu)$ by $P_{\mathcal{C}}(\nu)$ and replacing $R_{\mathcal{P}}$ by I. Observe that the replacement of $R_{\mathcal{P}}$ by I quantifies the difference between $P_{\mathcal{P}}^1(\nu)$ and $P_{\mathcal{P}}^2(\nu)$, and hence $P_{\mathcal{P}}^1(\nu)$ is always a "better" approximation to $P_{\mathcal{P}}(\nu)$ than $P_{\mathcal{P}}^2(\nu)$. However, regardless of the choice of preconditioner, the dominant cost is the solve with the matrix $P_{\mathcal{C}}(\nu)$ of (4.4). Note that $A_{\mathcal{B}}$ does not appear in $P_{\mathcal{P}}^2(\nu)$, which may make $P_{\mathcal{P}}^2(\nu)$ the more attractive preconditioner when it is expensive to form A, e.g., in PDE-constrained optimization [5].

It remains to establish the theoretical properties of the preconditioners $P_{\mathcal{P}}^1(\nu)$ and $P_{\mathcal{P}}^2(\nu)$. The next result shows that, asymptotically, the eigenvalues of $P(\nu)^{-1}B(\nu)$ and $P_{\mathcal{P}}^1(\nu)^{-1}B(\nu)$ are identical.

THEOREM 4.1 (properties of the preconditioner $P_{\mathcal{P}}^1(\nu)$). *Let $B(\nu)$ and $P_{\mathcal{P}}^1(\nu)$ be as defined in Definition* 2.1 *and* (4.5)*, respectively. In addition, assume that assumptions* (P'_1)–(P'_3) *hold. Then $P_{\mathcal{P}}^1(\nu)$ is positive definite for all $\nu > 0$. Moreover, the following properties hold for all $\nu \neq 0$:*

- (a) *The spectrum of $P_{\mathcal{P}}^1(\nu)^{-1}B(\nu)$ is independent of ν and consists of m unit eigenvalues and the n eigenvalues of $(M + A_{\mathcal{C}}^T D_{\mathcal{C}}^{-1} A_{\mathcal{C}})^{-1}(H + A^T D^{-1} A)$.*
- (b) *The matrix $P_{\mathcal{P}}^1(\nu)^{-1}B(\nu)$ has all eigenvalues of order $O(1)$ and at least $m + \mathrm{rank}(A_{\mathcal{S}})$ eigenvalues $1 + O(\mu^{1/2})$, of which at least m are exactly one.*

(c) *If $P(\nu)$ is the preconditioner of Definition* 3.3*, then* $\operatorname{eig}(P_{\mathcal{P}}^1(\nu)^{-1}B(\nu)) = \operatorname{eig}(P(\nu)^{-1}B(\nu)) + O(\mu)$.

Proof. Proposition 2.3 implies that the preconditioner $P_{\mathcal{P}}^1(\nu)$ is positive definite for all $\nu > 0$. For the remainder of the proof it will be assumed that ν is nonzero. It follows that $\operatorname{eig}(P_{\mathcal{P}}^1(\nu)^{-1}B(\nu)) = \operatorname{eig}(P_{\mathcal{P}}^2(\nu)^{-1}R_{\mathcal{P}}^{-1}B(\nu)R_{\mathcal{P}}^{-T})$, where

$$R_{\mathcal{P}}^{-1}B(\nu)R_{\mathcal{P}}^{-T} = \begin{pmatrix} H + A_{\mathcal{B}}^T D_{\mathcal{B}}^{-1} A_{\mathcal{B}} + (1+\nu) A_{\mathcal{C}}^T D_{\mathcal{C}}^{-1} A_{\mathcal{C}} & \nu A_{\mathcal{C}}^T & \\ \nu A_{\mathcal{C}} & \nu D_{\mathcal{C}} & \\ & & \nu D_{\mathcal{B}} \end{pmatrix}. \tag{4.7}$$

By successively replacing H by $H + A_{\mathcal{B}}^T D_{\mathcal{B}}^{-1} A_{\mathcal{B}}$, A by $A_{\mathcal{C}}$, and D by $D_{\mathcal{C}}$ in Lemma 3.4, a combination of (4.6) and (4.7) gives

$$P_{\mathcal{P}}^2(\nu)^{-1}R_{\mathcal{P}}^{-1}B(\nu)R_{\mathcal{P}}^{-T} = \begin{pmatrix} S_{\mathcal{C}} & & \\ T_{\mathcal{C}} & I & \\ & & I \end{pmatrix}, \tag{4.8}$$

where the matrices $S_{\mathcal{C}}$ and $T_{\mathcal{C}}$ are given by

$$S_{\mathcal{C}} = (M + A_{\mathcal{C}}^T D_{\mathcal{C}}^{-1} A_{\mathcal{C}})^{-1}(H + A^T D^{-1} A), \tag{4.9a}$$

$$T_{\mathcal{C}} = D_{\mathcal{C}}^{-1} A_{\mathcal{C}} (M + A_{\mathcal{C}}^T D_{\mathcal{C}}^{-1} A_{\mathcal{C}})^{-1}(M - H - A_{\mathcal{B}}^T D_{\mathcal{B}}^{-1} A_{\mathcal{B}}). \tag{4.9b}$$

The identity (4.8) implies that the spectrum of $P_{\mathcal{P}}^1(\nu)^{-1}B(\nu)$ consists of the eigenvalues of $S_{\mathcal{C}}$ and m unit eigenvalues, which proves part (a). Since $S_{\mathcal{C}}$ is independent of ν, the spectrum of $P_{\mathcal{P}}^1(\nu)^{-1}B(\nu)$ is also independent of ν. Lemma 3.5 implies that $S_{\mathcal{C}}$ has at least $\operatorname{rank}(A_{\mathcal{S}})$ eigenvalues that are $1 + O(\mu^{1/2})$, which establishes part (b).

To establish part (c), we need to estimate the difference between the eigenvalues of $S_{\mathcal{C}}$ and S, where S is given by (3.3a). This can be done using Lemma A.2 of the appendix. We may write

$$S_{\mathcal{C}} = I + (M + A_{\mathcal{C}}^T D_{\mathcal{C}}^{-1} A_{\mathcal{C}})^{-1}(H - M) + (M + A_{\mathcal{C}}^T D_{\mathcal{C}}^{-1} A_{\mathcal{C}})^{-1} A_{\mathcal{B}}^T D_{\mathcal{B}}^{-1} A_{\mathcal{B}}, \tag{4.10a}$$

$$S = I + (M + A^T D^{-1} A)^{-1}(H - M). \tag{4.10b}$$

By assumption, matrix $M + A_{\mathcal{C}}^T D_{\mathcal{C}}^{-1} A_{\mathcal{C}}$ is positive definite with the smallest eigenvalue bounded away from zero, and $A_{\mathcal{B}}^T D_{\mathcal{B}}^{-1} A_{\mathcal{B}}$ is positive definite with $\|D_{\mathcal{B}}^{-1}\| = O(\mu)$. The identity (4.10a) implies that the matrix $(M + A_{\mathcal{C}}^T D_{\mathcal{C}}^{-1} A_{\mathcal{C}})^{1/2} S_{\mathcal{C}} (M + A_{\mathcal{C}}^T D_{\mathcal{C}}^{-1} A_{\mathcal{C}})^{-1/2}$ is symmetric and has the same eigenvalues as $S_{\mathcal{C}}$, and it follows from (4.10) that

$$\operatorname{eig}(S_{\mathcal{C}}) = 1 + \operatorname{eig}((M + A_{\mathcal{C}}^T D_{\mathcal{C}}^{-1} A_{\mathcal{C}})^{-1}(H - M)) + O(\mu), \tag{4.11a}$$

$$\operatorname{eig}(S) = 1 + \operatorname{eig}((M + A^T D^{-1} A)^{-1}(H - M)). \tag{4.11b}$$

If we define $M_1 = M + A_{\mathcal{C}}^T D_{\mathcal{C}}^{-1} A_{\mathcal{C}}$, $M_2 = A_{\mathcal{B}}^T D_{\mathcal{B}}^{-1} A_{\mathcal{B}}$, and $M_3 = H - M$, then (4.11) gives $\operatorname{eig}(S_{\mathcal{C}}) = 1 + \operatorname{eig}(M_1^{-1} M_3) + O(\mu)$ and $\operatorname{eig}(S) = 1 + \operatorname{eig}((M_1 + M_2)^{-1} M_3)$. Lemma A.2 in conjunction with assumptions (P_1')–(P_3') gives the desired result. ∎

Next we establish that $P_{\mathcal{P}}^2(\nu)$ has the same asymptotic behavior as $P(\nu)$ and $P_{\mathcal{P}}^1(\nu)$. For $P_{\mathcal{P}}^2(\nu)$ it is assumed that $\nu > 0$, which ensures that the eigenvalues of $P_{\mathcal{P}}^2(\nu)^{-1}B(\nu)$ are real. The preconditioner $P_{\mathcal{P}}^2(\nu)$ is less expensive to apply than $P_{\mathcal{P}}^1(\nu)$, but the number of unit eigenvalues of the preconditioned matrix decreases from m to $m - \operatorname{rank}(A_{\mathcal{B}})$ because $A_{\mathcal{B}}$ does not appear in $P_{\mathcal{P}}^2(\nu)$. However, as the next theorem shows, for $\nu > 0$, $P_{\mathcal{P}}^2(\nu)$ behaves almost as well as $P_{\mathcal{P}}^1(\nu)$ in the sense that the eigenvalues of $P_{\mathcal{P}}^2(\nu)^{-1}B(\nu)$ differ from the eigenvalues of $P(\nu)^{-1}B(\nu)$ by $O(\mu^{1/2})$.

THEOREM 4.2 (properties of the preconditioner $P_{\mathcal{P}}^2(\nu)$). *Let $B(\nu)$ and $P_{\mathcal{P}}^2(\nu)$ be as defined in Definition* 2.1 *and* (4.6)*, respectively. In addition, assume that assumptions* (P_1')–(P_3') *hold. Then the following properties hold for all $\nu > 0$:*

(a) *$P_{\mathcal{P}}^2(\nu)$ is positive definite.*
(b) *The matrix $P_{\mathcal{P}}^2(\nu)^{-1}B(\nu)$ has all eigenvalues of order $O(1)$ and at least $m + \operatorname{rank}(A_{\mathcal{S}})$ eigenvalues $1 + O(\mu^{1/2})$, of which at least $m - \operatorname{rank}(A_{\mathcal{B}})$ are exactly one.*
(c) *If $P(\nu)$ is the preconditioner of Definition* 3.3*, then* $\operatorname{eig}(P_{\mathcal{P}}^2(\nu)^{-1}B(\nu)) = \operatorname{eig}(P(\nu)^{-1}B(\nu)) + O(\mu^{1/2})$.

Proof. The positive definiteness of $P_{\mathcal{P}}^2(\nu)$ for $\nu > 0$ follows from Proposition 2.3. For the remainder of the proof, assume that $\nu > 0$. Then, since $P_{\mathcal{P}}^2(\nu)$ is positive definite, the identity $\operatorname{eig}(P_{\mathcal{P}}^2(\nu)^{-1/2}B(\nu)P_{\mathcal{P}}^2(\nu)^{-1/2}) = \operatorname{eig}(P_{\mathcal{P}}^2(\nu)^{-1}B(\nu))$ ensures that $P_{\mathcal{P}}^2(\nu)^{-1}B(\nu)$ has real eigenvalues. Analogous to the proof of Theorem 4.1, by successively replacing the matrix H by $H + (1+\nu)A_{\mathcal{B}}^T D_{\mathcal{B}}^{-1} A_{\mathcal{B}}$, A by $A_{\mathcal{C}}$, and D by $D_{\mathcal{C}}$ in Lemma 3.4, and by using a combination of Proposition 2.3 and (4.6), we find that

$$P_{\mathcal{P}}^2(\nu)^{-1}B(\nu) = \begin{pmatrix} S_{\mathcal{C}} & & \\ T_{\mathcal{C}} & I & \\ & & I \end{pmatrix} + \begin{pmatrix} U & & X \\ V & & Y \\ W & & \end{pmatrix}, \tag{4.12}$$

where $S_{\mathcal{C}}$ and $T_{\mathcal{C}}$ are given by (4.9),

$$U = \nu(M + A_{\mathcal{C}}^T D_{\mathcal{C}}^{-1} A_{\mathcal{C}})^{-1} A_{\mathcal{B}}^T D_{\mathcal{B}}^{-1} A_{\mathcal{B}}, \tag{4.13a}$$

$$V = -\nu D_{\mathcal{C}}^{-1} A_{\mathcal{C}} (M + A_{\mathcal{C}}^T D_{\mathcal{C}}^{-1} A_{\mathcal{C}})^{-1} A_{\mathcal{B}}^T D_{\mathcal{B}}^{-1} A_{\mathcal{B}}, \tag{4.13b}$$

$$W = D_{\mathcal{B}}^{-1} A_{\mathcal{B}}, \tag{4.13c}$$

$$X = \nu(M + A_{\mathcal{C}}^T D_{\mathcal{C}}^{-1} A_{\mathcal{C}})^{-1} A_{\mathcal{B}}^T, \tag{4.13d}$$

$$Y = -\nu D_{\mathcal{C}}^{-1} A_{\mathcal{C}} (M + A_{\mathcal{C}}^T D_{\mathcal{C}}^{-1} A_{\mathcal{C}})^{-1} A_{\mathcal{B}}^T. \tag{4.13e}$$

It follows from (4.12) that $P_{\mathcal{P}}^2(\nu)^{-1}B(\nu)$ contains $m - m_{\mathcal{B}}$ columns from the identity matrix; hence it has at least $m - m_{\mathcal{B}}$ unit eigenvalues. The remaining eigenvalues are those of the matrix N given by

$$N = \begin{pmatrix} S_{\mathcal{C}} + U & X \\ W & I \end{pmatrix} = \begin{pmatrix} S_{\mathcal{C}} + \nu S_M^{-1} A_{\mathcal{B}}^T D_{\mathcal{B}}^{-1} A_{\mathcal{B}} & \nu S_M^{-1} A_{\mathcal{B}}^T \\ D_{\mathcal{B}}^{-1} A_{\mathcal{B}} & I \end{pmatrix}, \tag{4.14}$$

with $S_M = M + A_{\mathcal{C}}^T D_{\mathcal{C}}^{-1} A_{\mathcal{C}}$. Observe that (4.14) implies that any nonzero vector x such that $A_{\mathcal{B}}^T x = 0$ induces an eigenvector corresponding to a unit eigenvalue of N. Hence, N has at least $m_{\mathcal{B}} - \operatorname{rank}(A_{\mathcal{B}})$ unit eigenvalues. Further, let $\widetilde{Q}$ be defined by

$$\widetilde{Q} = \begin{pmatrix} S_M^{1/2} & \\ & \nu^{1/2} D_{\mathcal{B}}^{1/2} \end{pmatrix}.$$

Then N and $\widetilde{Q}N\widetilde{Q}^{-1}$ have identical eigenvalues, and $\widetilde{Q}N\widetilde{Q}^{-1}$ is given by

$$\begin{aligned} \widetilde{Q}N\widetilde{Q}^{-1} &= \begin{pmatrix} S_M^{1/2} S_{\mathcal{C}} S_M^{-1/2} + \nu S_M^{-1/2} A_{\mathcal{B}}^T D_{\mathcal{B}}^{-1} A_{\mathcal{B}} S_M^{-1/2} & \nu^{1/2} S_M^{-1/2} A_{\mathcal{B}}^T D_{\mathcal{B}}^{-1/2} \\ \nu^{1/2} D_{\mathcal{B}}^{-1/2} A_{\mathcal{B}} S_M^{-1/2} & I \end{pmatrix} \\ &= \begin{pmatrix} S_M^{1/2} S_{\mathcal{C}} S_M^{-1/2} & \\ & I \end{pmatrix} + \begin{pmatrix} \nu S_M^{-1/2} A_{\mathcal{B}}^T D_{\mathcal{B}}^{-1} A_{\mathcal{B}} S_M^{-1/2} & \nu^{1/2} S_M^{-1/2} A_{\mathcal{B}}^T D_{\mathcal{B}}^{-1/2} \\ \nu^{1/2} D_{\mathcal{B}}^{-1/2} A_{\mathcal{B}} S_M^{-1/2} & \end{pmatrix}. \end{aligned}$$

Note that $S_M^{1/2} S_{\mathcal{C}} S_M^{-1/2}$ is symmetric, and hence $\widetilde{Q} N \widetilde{Q}^{-1}$ is symmetric. In addition, from our assumptions, it follows that $\|\nu^{1/2} S_M^{-1/2} A_{\mathcal{B}}^T D_{\mathcal{B}}^{-1/2}\| = O(\mu^{1/2})$ and $\|\nu S_M^{-1/2} A_{\mathcal{B}}^T D_{\mathcal{B}}^{-1} A_{\mathcal{B}} S_M^{-1/2}\| = O(\mu)$. Hence, $\widetilde{Q} N \widetilde{Q}^{-1}$ has $m_{\mathcal{B}}$ eigenvalues that differ by $O(\mu^{1/2})$ from unity, and n eigenvalues that differ by $O(\mu^{1/2})$ from the eigenvalues of the matrix $S_M^{1/2} S_{\mathcal{C}} S_M^{-1/2}$. In addition, the eigenvalues of $S_M^{1/2} S_{\mathcal{C}} S_M^{-1/2}$ and $S_{\mathcal{C}}$ are identical. Consequently, it follows that the spectrum of $P_{\mathcal{P}}^2(\nu)^{-1} B(\nu)$ consists of $m - \operatorname{rank}(A_{\mathcal{B}})$ unit eigenvalues, $\operatorname{rank}(A_{\mathcal{B}})$ eigenvalues that are $1 + O(\mu^{1/2})$, and n eigenvalues that differ by $O(\mu^{1/2})$ from the eigenvalues of $S_{\mathcal{C}}$. Theorem 4.1 now shows that $\operatorname{eig}(P_{\mathcal{P}}^2(\nu)^{-1} B(\nu)) = \operatorname{eig}(P_{\mathcal{P}}^1(\nu)^{-1} B(\nu)) + O(\mu^{1/2})$, which gives the required result, since $O(\mu^{1/2})$ dominates $O(\mu)$. In particular, Theorem 4.1 implies that $P_{\mathcal{P}}^2(\nu)^{-1} B(\nu)$ has at least $m + \operatorname{rank}(A_{\mathcal{S}})$ eigenvalues that are $1 + O(\mu^{1/2})$. ∎

We conclude that it is possible to construct appropriate constraint preconditioners based on solving the smaller system (4.4). Moreover, the matrix $P_{\mathcal{C}}(\nu)$ of (4.4) has exactly the same structure as $P(\nu)$. The difference is that the number of rows and columns in the preconditioner has been reduced from $n + m$ to $n + m - m_{\mathcal{B}}$. Hence, all the previous analysis applies. For our example with 100 variables and 100,000 inequality constraints, a matrix of dimension 150 would need to be factored instead of a matrix of dimension 100,100.

As shown above, the partition of the row indices into $\mathcal{B}$ and its complement $\mathcal{C}$ provides active-set preconditioners $P_{\mathcal{P}}^1(\nu)$ and $P_{\mathcal{P}}^2(\nu)$ that are asymptotically equivalent to $P(\nu)$. If strict complementarity holds, then $m_{\mathcal{M}} = 0$ and the division into large and small elements is straightforward. If strict complementarity does not hold, then $m_{\mathcal{M}} > 0$ and we have chosen to append $\mathcal{M}$ to $\mathcal{S}$. Analogous preconditioners may be constructed by first identifying $\mathcal{S}$ and then forming the complementary set $\bar{\mathcal{S}}$, which is the set obtained by appending $\mathcal{M}$ to $\mathcal{B}$. The resulting KKT system analogous to (4.4) would have smaller dimension because $\mathcal{C}$ is replaced by $\mathcal{S}$. However, the resulting preconditioners would not be asymptotically equivalent in general. For $P_{\mathcal{P}}^1(\nu)$, the $1 + O(\mu^{1/2})$ cluster of eigenvalues would be the same as for $P(\nu)$, but the eigenvalues resulting from M would differ by an $O(1)$ term. The reason for this difference is that the norm of $D_{\bar{\mathcal{S}}}^{-1}$ would not be of order $O(\mu)$, but would include terms involving $m_{\mathcal{M}}$ eigenvalues of order one.

It should be emphasized that the choice of $\mathcal{C}$ and $\mathcal{B}$ affects only the efficiency of the active-set constraint preconditioners and not the definition of the linear equations that need to be solved. A poorly chosen partition may adversely affect the quality of the preconditioner, but not the solution of the linear equations. The partition analyzed here provides the largest $\mathcal{B}$ for which we can guarantee that the preconditioners $P_{\mathcal{P}}^1(\nu)$ and $P_{\mathcal{P}}^2(\nu)$ are asymptotically equivalent to $P(\nu)$ for $\nu > 0$. If elements are excluded from $\mathcal{B}$, then $P_{\mathcal{P}}^1(\nu)$ and $P_{\mathcal{P}}^2(\nu)$ become "better" approximations to $P(\nu)$, and the asymptotic performance is unchanged. However, this increases the dimension of the KKT system (4.4). As noted above, if $\mathcal{B}$ is chosen too large, in the sense that diagonal elements of D are included in $D_{\mathcal{B}}$ that are not $\Omega(1/\mu)$, then the quality of the active-set preconditioners can be expected to deteriorate. Hence, it is not essential that $\mathcal{B}$ is estimated correctly, but it is essential that $D_{\mathcal{B}}$ contains only large elements.

5. On semidefinite diagonal matrices. Up to this point we have assumed that the matrix D in the $(2,2)$ block of the KKT system is positive definite. In the general case, the last block of equations in the KKT system has the form

$$Ax_1 + Gx_2 = b_2, \tag{5.1}$$

where G is a diagonal matrix with positive and zero entries. If all the constraints of the optimization problem are nonlinear, it is always possible to formulate the interior method so that G is positive definite. For inequality constraints, standard formulations give positive elements in G that are of the order of the perturbation parameter μ (see, e.g., Vanderbei and Carpenter [46] and Forsgren and Gill [16]). Typically, zero elements of G are associated with linearized equality constraints, where the corresponding subset of equations (5.1) are the Newton equations for a zero of the constraint residual. An alternative to direct constraint linearization is to impose equality constraints approximately via a quadratic penalty function. It can be shown that this approach gives a positive element in G of the order of $\bar{\mu}$, where $\bar{\mu}$ is the inverse of the penalty parameter (see, e.g., Gould [31] and Forsgren and Gill [16]). The parameter $\bar{\mu}$ may be allowed to vary with μ, or may be fixed at some small value (see, e.g., Gill et al. [26] and Saunders and Tomlin [42]). Fixing $\bar{\mu}$ defines a *regularization* of the problem, which allows the formulation of methods that do not require an assumption on the rank of the equality constraint Jacobian. (For more details on the use of regularization in interior methods, see Gill et al. [24], Vanderbei and Shanno [47], and Altman and Gondzio [1].)

However, it may not always be beneficial to regularize *linear* constraints. Regularization in this context is less crucial because reliable techniques exist for discarding dependent equality constraints. Moreover, interior methods can be defined so that every iterate satisfies the linear equality constraints (see below). With an appropriate choice of constraints, this feature can be used to guarantee that the nonlinear functions and their derivatives are well defined at all points generated by the interior method.

In order to consider KKT systems with a semidefinite $(2,2)$ block, we assume that the variables and equations are preordered to give a system $Bx = b$ such that

$$\begin{pmatrix} H & -A^T & -F^T \\ -A & -D & \\ -F & & \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} b_1 \\ -b_2 \\ -b_3 \end{pmatrix}, \tag{5.2}$$

where D is positive definite. Note that we cannot compute the condensed or doubly augmented system for these equation because of the zero block. In this case, B has correct inertia if $N^T(H + A^T D^{-1} A)N$ is positive definite, where the columns of N form a basis for the null space of F (see Forsgren [18]).

The KKT system (5.2) may be solved using a projection technique similar to that described in section 3. First, an initial point y is found with first n components forming a vector y_1 such that $Fy_1 = b_3$. This vector may be computed in various ways—e.g., by computing an LU factorization of F^T (see, e.g., Gill, Murray, and Saunders [27]), or by solving a system for the preconditioning matrix associated with (5.2), where H is replaced by a suitable approximation M (see Gould, Hribar, and Nocedal [29]). Once y is known, the PCG method may be used to solve

$$\begin{pmatrix} H & -A^T & -F^T \\ -A & -D & \\ -F & & \end{pmatrix} \begin{pmatrix} \widehat{x}_1 \\ \widehat{x}_2 \\ \widehat{x}_3 \end{pmatrix} = \begin{pmatrix} \widehat{b}_1 \\ -\widehat{b}_2 \\ 0 \end{pmatrix}, \tag{5.3}$$

with $\widehat{b}_1 = b_1 - Hy_1 + A^T y_2 + F^T y_3$ and $\widehat{b}_2 = b_2 - Ay_1 - Dy_2$. The required solution is then $x = y + \widehat{x}$. Analogous to the situation when G is positive definite, we may

embed (5.3) into the parameterized system of linear equations

$$\begin{pmatrix} H+(1+\nu)A^TD^{-1}A & \nu A^T & -F^T \\ \nu A & \nu D & \\ -F & & \end{pmatrix} \begin{pmatrix} \widehat{x}_1 \\ \widehat{x}_2 \\ \widehat{x}_3 \end{pmatrix} = \begin{pmatrix} \widehat{b}_1+(1+\nu)A^TD^{-1}\widehat{b}_2 \\ \nu\widehat{b}_2 \\ 0 \end{pmatrix}, \tag{5.4}$$

with $\widehat{b}_1 = b_1 - Hy_1 + A^Ty_2 + F^Ty_3$ and $\widehat{b}_2 = b_2 - Ay_1 - Dy_2$. If the zero elements of G are associated with linear constraints, and the system (5.3) is solved exactly, it suffices to compute the special step y only once, when solving the first system. Then, provided that the constraint preconditioner is applied exactly at every PCG step, the right-hand side of (5.3) will remain zero for all subsequent iterations.

The linear equations (5.2), (5.3), and (5.4) do not require N. If the preconditioner cannot be applied exactly, then it is necessary to use an alternative method based on computing products of the form N^Tv and Nu. (Gill, Murray, and Saunders [27] describe how these products may be computed in a numerically stable way without needing to store N explicitly.) The requirement that $F\widehat{x}_1 = 0$ implies that $\widehat{x}_1$ can be written as $\widehat{x}_1 = N\hat{p}_1$. Substituting this expression in (5.3) gives the reduced KKT system

$$\begin{pmatrix} N^THN & -(AN)^T \\ -AN & -D \end{pmatrix} \begin{pmatrix} \hat{p}_1 \\ \hat{p}_2 \end{pmatrix} = \begin{pmatrix} N^T(b_1 - Hy_1 + A^Ty_2) \\ -b_2 + Ay_1 + Dy_2 \end{pmatrix}, \tag{5.5}$$

from which we can define $\widehat{x}_1 = N\hat{p}_1$ and $\widehat{x}_2 = \hat{p}_2$. This system has a nonsingular $(2,2)$ block and has correct inertia if (5.2) has correct inertia. Moreover, the iterates define exact projections regardless of the accuracy of the solves with the preconditioner. Hence, all the conditions needed for the application of the PCG method proposed in section 3 apply. The systems (5.3) and (5.5) are mathematically equivalent, which implies that we may apply the analysis of section 3 directly to both (5.3) and (5.4).

6. Some numerical examples. To illustrate the numerical performance of the proposed preconditioners, a PCG method was applied to a collection of illustrative large sparse KKT systems. The test matrices were generated from a number of realistic KKT systems arising in the context of primal-dual interior methods. We conclude with some randomly generated problems that illustrate some of the properties of the preconditioned matrices.

6.1. Examples from the COPS test set. First we describe some numerical results obtained on linear equations arising in a primal-dual interior method applied to optimization problems from the COPS 3.0 test collection [6, 9, 10] implemented in the AMPL modeling language [2, 19].

The equations are analogous to those generated by an interior-point method with barrier parameter μ. The data for the test matrices was generated using a primal-dual trust-region method (see, e.g., [16, 20, 32]) applied to eight problems, *Camshape*, *Channel*, *Gasoil*, *Marine*, *Methanol*, *Pinene*, *Polygon*, and *Tetra*, from the COPS 3.0 test collection [6, 8, 9, 10]. The interior-point method requires the solution of systems with a KKT matrix of the form

$$\begin{pmatrix} H & -J^T \\ -J & -\Gamma \end{pmatrix}, \tag{6.1}$$

where H is the $n \times n$ Hessian of the Lagrangian, J is the $m \times n$ Jacobian matrix of constraint gradients, and Γ is a positive-definite diagonal with some large and small

TABLE 6.1
Dimensions of the AMPL versions of the COPS problems.

Problem	n	m	$\|H - M\|$	$\sigma_k(J)$
Camshape	1200	3600	2.6e+0	1.3e-5
Channel	6398	6398	1.1e+2	4.1e-5
Gasoil	4001	4001	1.1e+1	0.0e+0
Marine	6415	6407	3.5e+1	0.0e+0
Methanol	4802	4802	3.1e+0	2.8e-3
Pinene	8000	8000	9.2e+3	6.8e-8
Polygon	398	20496	2.4e+2	0.0e+0
Tetra	1200	4254	2.7e+1	0.0e+0

elements. These systems have the same structure as the generic system (1.2). The dimensions of the eight problems are given in Table 6.1. The optimization problems in the COPS collection have a mixture of general nonlinear constraints and simple upper and lower bounds on the variables. The simple bounds lead to unit rows in J, and it is customary to define a smaller KKT system in which the unit rows and columns are eliminated. However, in the numerical experiments, the unit rows were included in order to more accurately illustrate the results of Theorems 3.6, 4.1, and 4.2. Hence the value of m also includes the bound constraints. The final column gives the kth largest singular value of J, where $k = \min\{m, n\}$.

For each of the eight featured COPS problems, matrices H, A, D and the right-hand side were generated from the matrices H, J, Γ and the right-hand side at a snapshot taken at iteration 30 of the interior method. For each problem snapshot, five systems of equations were generated by specifying five matrices D with entries parameterized by a scalar $\mu = 10^{-\ell}$ for $\ell = \{1, 2, 4, 6, 8\}$. For each value of μ, the matrices A and D were generated from J and Γ using the MATLAB code fragment

```
[D, ind] = sort(Gamma); % sort the diagonals of Gamma
A  = J(ind,:);          % reorder the rows of J
k  = min([m,n]);        %
if k < m, D(k+1:end) = max(D(k+1:end), 1/mu); end
D(1:k) = min(D(1:k),mu);
```

This choice of D implicitly defines a sequence of systems associated with a vertex solution of the underlying optimization problem for which strict complementarity holds. This was done deliberately to minimize the effect of the matrix M on the efficiency of the preconditioner (see the definition of $P(\nu)$ in (4.1)). Asymptotically, the matrix M defines the efficiency of the preconditioner within the null space of the matrix of active constraints (see, e.g., Dollar et al. [11]). In our analysis we have focused on the part of the preconditioner associated with the constraint part of the KKT system. The formulation and analysis of effective choices for M are beyond the scope of this paper. (For some possible approaches, see, e.g., [22, 23].) In the experiments reported here, M was a diagonal matrix with entries $M_{jj} = \max\left(|H_{jj}|, \delta\right)$, where $\delta = 10^{-1}$.

Figure 6.1 depicts the number of PCG iterations required to solve the resulting 8 sets of 5 systems of linear equations. The bar charts give the PCG iterations for the condensed system (top) and doubly augmented system (bottom). The MATLAB version of SYMMLQ [39] was used as the PCG solver. The symmetric indefinite solver MA27 was used to factor the constraint preconditioner (see Duff and Reid [13]). The

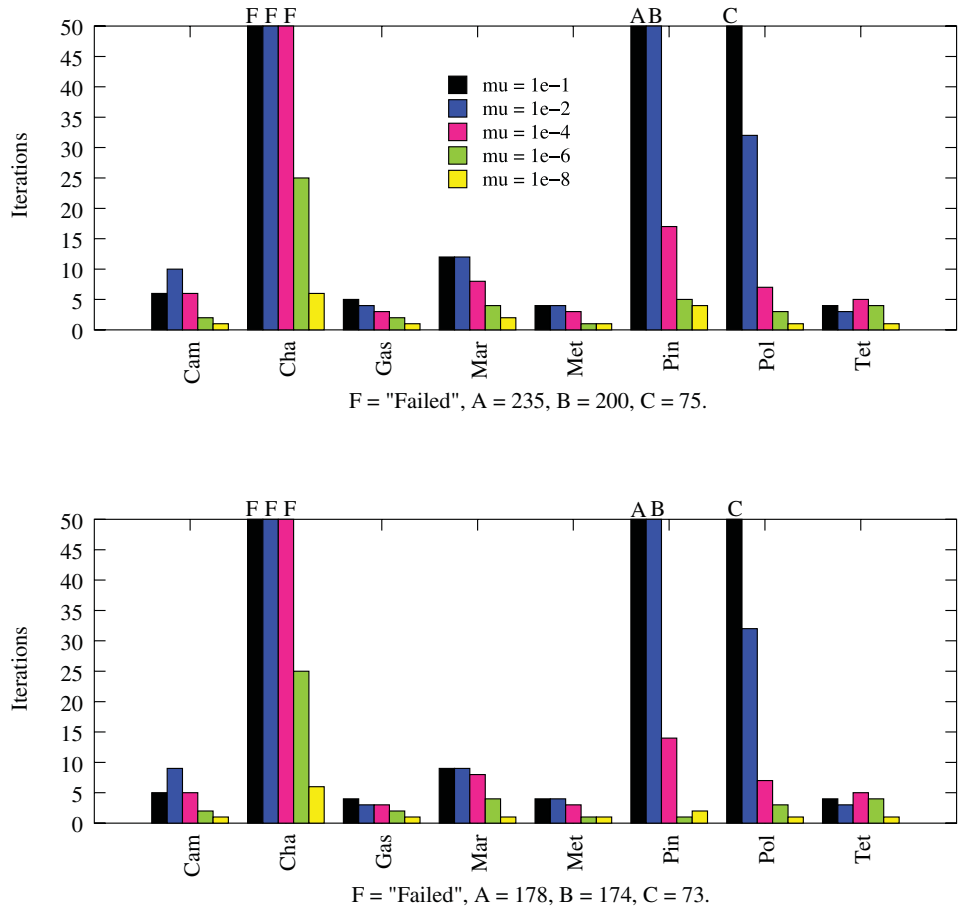


FIG. 6.1. *Results on the COPS problems. The graphs give the number of PCG iterations for* $P(0)$ *(above) and* $P(1)$ *(below). The vertical axis is limited to* 50 *iterations. The number of iterations needed for each off-scale case is given by the appropriate key code "A", "B", or "C".*

value of 10^{-6} was used for the SYMMLQ relative convergence tolerance. For problem *Channel*, with the three larger values of μ, PCG did not converge within the pre-assigned limit of 10^6 iterations (a more sophisticated choice of M is needed in this case). Note the similar number of PCG iterations needed to solve the condensed system and doubly augmented system.

Figure 6.2 gives the number of PCG iterations for the active-set preconditioners on the COPS problems *Camshape*, *Polygon*, and *Tetra*. These problems have significantly more constraints than variables and provide good examples on which to test the active-set preconditioners. In order to illustrate the behavior of the active-set preconditioner, we scale D so that exactly n elements are less than μ and the remaining elements are greater than $1/\mu$. We emphasize that the motivation for manipulating D in this way is to illustrate the effect of changing μ for fixed H and J.

```
mS = min([n,m,k]);
JS = J(1:mS,:);     JB = J(mS+1:end,:);
DS = D(1:mS);       DB = D(mS+1:end);
DS = spdiags(DS,0,mS,mS);
```

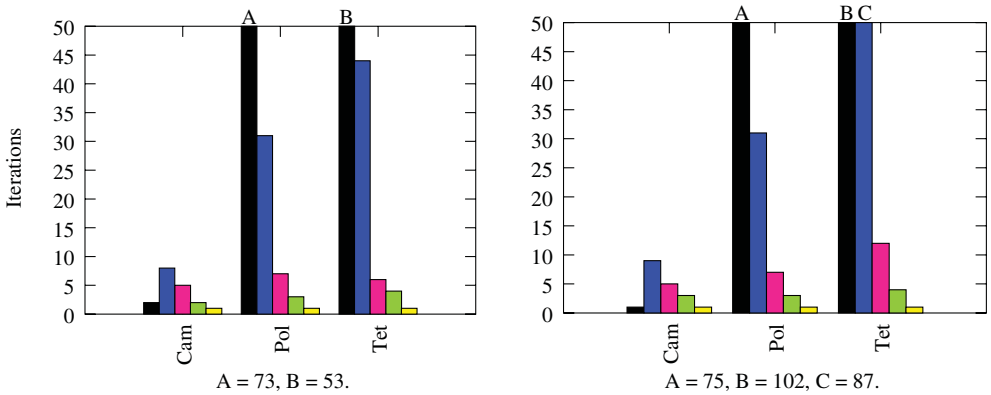


FIG. 6.2. *COPS problems: PCG iterations for $P^1_{\mathcal{P}}(1)$ (left) and $P^2_{\mathcal{P}}(1)$ (right). The vertical axis is limited to 50 iterations. The number of iterations needed for each off-scale case is given by the appropriate key code "A", "B", or "C".*

6.2. Results from randomly generated problems. Additional experiments were performed on randomly generated KKT systems. The purpose of these experiments was to illustrate the clustering of the eigenvalues of the preconditioned matrices associated with the doubly augmented system. The first set of experiments involved applying the preconditioners $P(1)$, $P^1_{\mathcal{P}}(1)$, and $P^2_{\mathcal{P}}(1)$ to randomly generated problems satisfying the assumptions of Theorems 4.1 and 4.2. Of particular interest is the "strict complementarity" assumption that every element of the diagonal D is either big or small. Given values $n = 400$, $m = 600$, and $\mu = 10^{-\ell}$ for $\ell = \{1, 2, 4, 6, 8\}$, matrices H, $A_{\mathcal{S}}$, $A_{\mathcal{B}}$, $D_{\mathcal{S}}$, and $D_{\mathcal{B}}$ were generated using the MATLAB code fragment

```
mS = 100; d  = 10^(-2);
H  = sprandsym(n,d);          JS = sprand(mS,n,d,0.1);  JB = sprand(m-mS,n,d);
DS = diag(mu*ones(mS,1));     DB = diag((1/mu)*ones(m-mS,1));
```

Table 6.2 gives details of the eigenvalues of the preconditioned matrices associated with each of the preconditioners $P(1)$, $P^1_{\mathcal{P}}(1)$, and $P^2_{\mathcal{P}}(1)$, where the diagonal preconditioner M was defined as in the COPS examples of the previous section. In all these runs, the resulting KKT matrix satisfies $\|H - M\| = 4.95$ and $\sigma_n(J_{\mathcal{S}}) = 10^{-1}$. These linear systems would be typical for a primal-dual method applied to an optimization problem with 100 active constraints at a point satisfying a strict complementarity assumption. Theorems 3.6 and 4.1 predict that for the preconditioners $P(1)$ and $P^1_{\mathcal{P}}(1)$, 700 $(= m + \text{rank}(A_{\mathcal{S}}))$ eigenvalues of the preconditioned matrix will cluster close to unity, with 600 of these eigenvalues exactly equal to one. Theorem 4.2 predicts that as μ is reduced, $P^2_{\mathcal{P}}(1)$ also will give 700 eigenvalues close to one, whereas 200 $(= m - n)$ eigenvalues will be exactly one.

The last four columns of Table 6.2 illustrate the degree of clustering of the eigenvalues of the preconditioned matrix. Clustering is measured by means of the function $l(\theta)$ defined as follows. Given a matrix C with real eigenvalues, the function

$$l(\theta) = \text{card}\{\lambda \in \text{eig}(C) : |\lambda - 1| \leq \theta\}$$

gives the number of eigenvalues of C within distance θ of unity. Table 6.2 gives the values of $l(\theta)$ for the three preconditioned matrices $C = P(1)^{-1}B(1)$, $P^1_{\mathcal{P}}(1)^{-1}B(1)$, and $P^2_{\mathcal{P}}(1)^{-1}B(1)$. In this strict-complementarity case, we expect that the proposed preconditioners would asymptotically give a cluster of 700 unit eigenvalues. Note that

TABLE 6.2
Number of clustered eigenvalues of the preconditioned matrix. Randomly generated KKT systems with $n = 400$, $m = 600$, $m_{\mathcal{S}} = 100$, $m_{\mathcal{M}} = 0$, $m_{\mathcal{B}} = 500$, $(D_{\mathcal{S}})_{ii} = \mu$, and $(D_{\mathcal{B}})_{ii} = 1/\mu$.

	μ	$l(10^{-8})$	$l(10^{-6})$	$l(10^{-4})$	$l(10^{-2})$
P	10^{-1}	600	600	600	601
	10^{-2}	600	600	600	612
	10^{-4}	600	600	612	675
	10^{-6}	600	612	673	700
	10^{-8}	606	673	700	700
$P^1_{\mathcal{P}}$	10^{-1}	600	600	600	602
	10^{-2}	600	600	600	611
	10^{-4}	600	600	612	675
	10^{-6}	600	612	673	700
	10^{-8}	593	673	700	700
$P^2_{\mathcal{P}}$	10^{-1}	200	200	200	425
	10^{-2}	200	200	215	567
	10^{-4}	200	215	566	673
	10^{-6}	215	566	672	700
	10^{-8}	546	672	700	700

TABLE 6.3
Number of clustered eigenvalues for the preconditioned matrix. Randomly generated KKT systems with $n = 400$, $m = 600$, $m_{\mathcal{S}} = 75$, $m_{\mathcal{M}} = 25$, $m_{\mathcal{B}} = 500$, $(D_{\mathcal{S}})_{ii} = \mu$, $(D_{\mathcal{M}})_{ii} = 1$, and $(D_{\mathcal{B}})_{ii} = 1/\mu$.

	μ	$l(10^{-8})$	$l(10^{-6})$	$l(10^{-4})$	$l(10^{-2})$
P	10^{-1}	600	600	600	601
	10^{-2}	600	600	600	609
	10^{-4}	600	600	609	654
	10^{-6}	600	609	653	675
	10^{-8}	604	653	675	675
$P^1_{\mathcal{P}}$	10^{-1}	600	600	600	602
	10^{-2}	600	600	600	609
	10^{-4}	600	600	609	654
	10^{-6}	600	609	653	675
	10^{-8}	591	653	675	675
$P^2_{\mathcal{P}}$	10^{-1}	200	200	200	425
	10^{-2}	200	200	215	563
	10^{-4}	200	215	563	653
	10^{-6}	215	563	653	675
	10^{-8}	544	653	675	675

for small values of μ, $P(1)$ and $P^1_{\mathcal{P}}(1)$ produce very similar numbers of eigenvalues close to unity. The preconditioner $P^2_{\mathcal{P}}(1)$ tends to give fewer accurate eigenvalues than $P(1)$ and $P^1_{\mathcal{P}}(1)$ for the larger values of μ, although the differences become less marked as μ is reduced.

Table 6.3 was generated with the same data used for Table 6.2, with the one exception that strict complementarity was assumed not to hold. As in Table 6.2, we simulate an optimization problem with 100 active constraints, but in this case we set $m_{\mathcal{S}} = 75$ and $m_{\mathcal{M}} = 25$. The corresponding diagonal elements of $D_{\mathcal{M}}$ were set at one.

In this non-strict-complementarity case, Theorems 3.6, 4.1, and 4.2 predict that the proposed preconditioners would asymptotically give a cluster of 675 ($= m + m_{\mathcal{S}}$) unit eigenvalues, which is reflected in the results. The performance of the preconditioners is very similar to that depicted in Table 6.2.

7. Summary and further research. A framework has been proposed for applying the PCG method to KKT systems of the form (1.1) that arise in interior methods for general nonconvex optimization. The proposed methods are based on applying the conjugate-gradient method to the doubly augmented system (1.4), which is positive definite if the underlying optimization problem satisfies the second-order sufficient conditions for optimality. An advantage of the doubly augmented system is that it is positive definite with respect to all of the variables.

We also have proposed a class of constraint preconditioners for the doubly augmented system. In particular, we have analyzed two ways of using an estimate of the active set to reduce the cost of applying the preconditioner when there are many inequality constraints. As the solution of the optimization problem is approached, these active-set preconditioners have theoretical performance comparable to constraint preconditioners that include all the constraints. An advantage of using preconditioning in conjunction with the doubly augmented system is that the linear equations used to apply the preconditioner need not be solved exactly. Future work will consider the analysis associated with these approximate preconditioners.

The focus of this paper has been on the formulation and analysis of *constraint* preconditioners. The next step is to consider "full" preconditioners based on estimating the matrix H in the $(1,1)$ block of the KKT equations. For example, a preconditioner may be defined using an incomplete inertia-controlling factorization of the KKT system (1.2). For more details on the inertia-controlling factorization for augmented systems in interior methods, see Forsgren and Gill [16] and Forsgren [18].

Appendix. Linear algebra. Here we review two results from linear algebra. The first gives the structure of the inverse of $B(\nu)$ and may be verified by direct multiplication.

LEMMA A.1. *Given a nonsingular symmetric matrix D, consider the matrix*

$$B(\nu) = \begin{pmatrix} H + (1+\nu)A^T D^{-1} A & \nu A^T \\ \nu A & \nu D \end{pmatrix},$$

where ν is a scalar. Then $B(\nu)$ may be factored in the form

$$B(\nu) = \begin{pmatrix} I & A^T D^{-1} \\ & I \end{pmatrix} \begin{pmatrix} H + A^T D^{-1} A & \\ & \nu D \end{pmatrix} \begin{pmatrix} I & \\ D^{-1} A & I \end{pmatrix}.$$

Moreover, if $H + A^T D^{-1} A$ is nonsingular and $\nu \neq 0$, then $B(\nu)$ is nonsingular, with inverse

$$\begin{aligned} B(\nu)^{-1} &= \begin{pmatrix} (H + A^T D^{-1} A)^{-1} & -(H + A^T D^{-1} A)^{-1} A^T D^{-1} \\ -D^{-1} A (H + A^T D^{-1} A)^{-1} & \frac{1}{\nu} D^{-1} + D^{-1} A (H + A^T D^{-1} A)^{-1} A^T D^{-1} \end{pmatrix} \\ &= \begin{pmatrix} I & \\ -D^{-1} A & I \end{pmatrix} \begin{pmatrix} (H + A^T D^{-1} A)^{-1} & \\ & \frac{1}{\nu} D^{-1} \end{pmatrix} \begin{pmatrix} I & -A^T D^{-1} \\ & I \end{pmatrix}. \end{aligned}$$

The second result provides bounds on the perturbation of the eigenvalues of $M_1^{-1} M_3$ when M_1 is perturbed by a positive-semidefinite matrix M_2.

LEMMA A.2. *Let M_1, M_2, and M_3 be $n \times n$ symmetric matrices with M_1 positive definite and M_2 positive semidefinite. Let $\{\lambda_i\}$ and $\{\widetilde{\lambda}_i\}$ denote the eigenvalues of $M_1^{-1}M_3$ and $(M_1+M_2)^{-1}M_3$, respectively. Assume that the $\{\lambda_i\}$ are ordered so that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, with the same ordering for $\{\widetilde{\lambda}_i\}$. Then*

$$0 \leq \lambda_i - \widetilde{\lambda}_i \leq \|M_1^{-1/2}M_2M_1^{-1/2}\|\widetilde{\lambda}_i \quad \textit{for all } i \textit{ such that } \lambda_i \geq 0,$$
$$0 \leq \widetilde{\lambda}_i - \lambda_i \leq -\|M_1^{-1/2}M_2M_1^{-1/2}\|\widetilde{\lambda}_i \quad \textit{for all } i \textit{ such that } \lambda_i < 0.$$

Proof. Let M, $\widetilde{M}$, and $\widetilde{I}$ be defined such that

$$M = M_1^{-1/2}M_3M_1^{-1/2}, \quad \widetilde{I} = I + M_1^{-1/2}M_2M_1^{-1/2}, \quad \text{and} \quad \widetilde{M} = \widetilde{I}^{-1/2}M\widetilde{I}^{-1/2}.$$

Then M, $\widetilde{I}$, and $\widetilde{M}$ are symmetric with $\widetilde{I}$ positive definite. A similarity transformation gives $\operatorname{eig}(M_1^{-1}M_3) = \operatorname{eig}(M_1^{-1/2}M_3M_1^{-1/2}) = \operatorname{eig}(M)$, which means that M has eigenvalues λ_i, $i = 1:n$. Similarly, we have

$$\begin{aligned}(M_1+M_2)^{-1}M_3 &= M_1^{-1/2}(I + M_1^{-1/2}M_2M_1^{-1/2})^{-1}M_1^{-1/2}M_3 \\ &= M_1^{-1/2}\widetilde{I}^{-1}MM_1^{1/2}. \end{aligned} \tag{A.1}$$

Successive similarity transformations of (A.1) with $M_1^{1/2}$ and $\widetilde{I}^{1/2}$ give

$$\operatorname{eig}\big((M_1+M_2)^{-1}M_3\big) = \operatorname{eig}(\widetilde{I}^{-1}M) = \operatorname{eig}(\widetilde{I}^{-1/2}M\widetilde{I}^{-1/2}) = \operatorname{eig}(\widetilde{M}),$$

which means that $\widetilde{M}$ has eigenvalues $\widetilde{\lambda}_i$, $i = 1:n$.

Now we relate λ_i to $\widetilde{\lambda}_i$. First, consider the case $\lambda_i \geq 0$. Since λ_i is an eigenvalue of M and $\widetilde{\lambda}_i$ is an eigenvalue of $\widetilde{M}$, with $\widetilde{M} = \widetilde{I}^{-1/2}M\widetilde{I}^{-1/2}$, the Courant–Fischer min-max theorem gives

$$\frac{1}{\|\widetilde{I}^{-1}\|}\widetilde{\lambda}_i \leq \lambda_i \leq \|\widetilde{I}\|\widetilde{\lambda}_i; \tag{A.2}$$

see, e.g., Golub and Van Loan [28, pp. 403–404]. Since $\widetilde{I} = I + M_1^{-1/2}M_2M_1^{-1/2}$ with M_2 positive semidefinite, it follows that $\|\widetilde{I}^{-1}\| \leq 1$ and $\|\widetilde{I}\| \leq 1 + \|M_1^{-1/2}M_2M_1^{-1/2}\|$. Hence, (A.2) gives

$$\widetilde{\lambda}_i \leq \lambda_i \leq (1 + \|M_1^{-1/2}M_2M_1^{-1/2}\|)\widetilde{\lambda}_i,$$

which is equivalent to the desired result when $\lambda_i \geq 0$,

$$0 \leq \lambda_i - \widetilde{\lambda}_i \leq \|M_1^{-1/2}M_2M_1^{-1/2}\|\widetilde{\lambda}_i.$$

For the case $\lambda_i < 0$, we apply the analysis above to the matrices $-\widetilde{M}$ and $-M$. Then, since $-\lambda_i$ is a positive eigenvalue of $-M$ and $-\widetilde{\lambda}_i$ is an eigenvalue of $-\widetilde{M}$, we conclude that

$$0 \leq -\lambda_i + \widetilde{\lambda}_i \leq -\|M_1^{-1/2}M_2M_1^{-1/2}\|\widetilde{\lambda}_i,$$

as required. ∎

Acknowledgments. The authors would like to thank Michael Gertz for many stimulating discussions on the theoretical and practical aspects of interior methods. The authors are also grateful to the referees for many constructive comments that significantly improved the presentation.

REFERENCES

[1] A. Altman and J. Gondzio, *Regularized symmetric indefinite systems in interior point methods for linear and quadratic optimization*, Optim. Methods Softw., 11/12 (1999), pp. 275–302.
[2] AMPL Home Page, http://www.ampl.com.
[3] M. Benzi, G. H. Golub, and J. Liesen, *Numerical solution of saddle point problems*, in Acta Numerica, 2005, Acta Numer. 14, Cambridge University Press, Cambridge, UK, 2005, pp. 1–137.
[4] L. Bergamaschi, J. Gondzio, and G. Zilli, *Preconditioning indefinite systems in interior point methods for optimization*, Comput. Optim. Appl., 28 (2004), pp. 149–171.
[5] G. Biros and O. Ghattas, *Inexactness issues in the Lagrange-Newton-Krylov-Schur method for PDE-constrained optimization*, in Large-Scale PDE-Constrained Optimization (Santa Fe, NM, 2001), Lect. Notes Comput. Sci. Eng. 30, Springer, Berlin, 2003, pp. 93–114.
[6] A. Bondarenko, D. Bortz, and J. J. Moré, *COPS: Large-Scale Nonlinearly Constrained Optimization Problems*, Technical Report ANL/MCS-TM-237, Mathematics and Computer Science division, Argonne National Laboratory, Argonne, IL, 1998 (revised October 1999).
[7] R. S. Dembo, S. C. Eisenstat, and T. Steihaug, *Inexact Newton methods*, SIAM J. Numer. Anal., 19 (1982), pp. 400–408.
[8] E. D. Dolan, J. J. Moré, and T. S. Munson, *Benchmarking Optimization Software with COPS* 3.0, Technical Memorandum ANL/MCS-TM-273, Argonne National Laboratory, Argonne, IL, 2004.
[9] E. D. Dolan and J. J. Moré, *Benchmarking Optimization Software with COPS*, Technical Memorandum ANL/MCS-TM-246, Argonne National Laboratory, Argonne, IL, 2000.
[10] E. D. Dolan and J. J. Moré, *Benchmarking optimization software with performance profiles*, Math. Program., 91 (2002), pp. 201–213.
[11] H. S. Dollar, N. I. M. Gould, W. H. A. Schilders, and A. J. Wathen, *On Iterative Methods and Implicit-Factorization Preconditioners for Regularized Saddle-Point Systems*, Report RAL-TR-2005-011, Rutherford Appleton Laboratory, Oxfordshire, UK, 2005.
[12] H. S. Dollar, *Extending Constraint Preconditioners for Saddle Point Problems*, Numerical Analysis Group Research Report NA-05/02, Oxford University Computing Laboratory, Oxford, UK, 2005.
[13] I. S. Duff and J. K. Reid, *MA*27: *A Set of Fortran Subroutines for Solving Sparse Symmetric Sets of Linear Equations*, Technical Report R-10533, Computer Science and Systems Division, AERE Harwell, Oxford, UK, 1982.
[14] H. C. Elman, V. E. Howle, J. N. Shadid, and R. S. Tuminaro, *A parallel block multi-level preconditioner for the* 3*d incompressible Navier-Stokes equations*, J. Comput. Phys., 187 (2003), pp. 504–523.
[15] A. Forsgren, P. E. Gill, and M. H. Wright, *Interior methods for nonlinear optimization*, SIAM Rev., 44 (2002), pp. 525–597.
[16] A. Forsgren and P. E. Gill, *Primal-dual interior methods for nonconvex nonlinear programming*, SIAM J. Optim., 8 (1998), pp. 1132–1152.
[17] A. Forsgren and W. Murray, *Newton methods for large-scale linear equality-constrained minimization*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 560–587.
[18] A. Forsgren, *Inertia-controlling factorizations for optimization algorithms*, Appl. Numer. Math., 43 (2002), pp. 91–107.
[19] R. Fourer, D. M. Gay, and B. W. Kernighan, *AMPL: A Modeling Language for Mathematical Programming*, Brooks/Cole–Thomson Learning, Pacific Grove, CA, 2003.
[20] E. M. Gertz and P. E. Gill, *A primal-dual trust-region algorithm for nonlinear programming*, Math. Program. Ser. B, 100 (2004), pp. 49–94.
[21] E. M. Gertz and J. D. Griffin, *Support Vector Machine Classifiers for Large Data Sets*, Technical Memorandum ANL/MCS-TM-289, Mathematics and Computer Science division, Argonne National Laboratory, Argonne, IL, 2005.
[22] O. Ghattas and J.-H. Bark, *Large-scale SQP methods for optimization of Navier-Stokes flows*, in Large-Scale Optimization with Applications, Part II (Minneapolis, MN, 1995), Springer, New York, 1997, pp. 247–270.

[23] O. Ghattas and C. E. Orozco, *A parallel reduced Hessian SQP method for shape optimization*, in Multidisciplinary Design Optimization (Hampton, VA, 1995), SIAM, Philadelphia, PA, 1997, pp. 133–152.
[24] P. E. Gill, W. Murray, D. B. Ponceleón, and M. A. Saunders, *Primal-Dual Methods for Linear Programming*, Report SOL 91-3, Department of Operations Research, Stanford University, Stanford, CA, 1991.
[25] P. E. Gill, W. Murray, D. B. Ponceleón, and M. A. Saunders, *Preconditioners for indefinite systems arising in optimization*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 292–311.
[26] P. E. Gill, W. Murray, D. B. Ponceleón, and M. A. Saunders, *Solving reduced KKT systems in barrier methods for linear programming*, in Numerical Analysis 1993 (Dundee, 1993), G. A. Watson and D. Griffiths, eds., Pitman Res. Notes Math. Ser. 303, Longman Sci. Tech., Harlow, UK, 1994, pp. 89–104.
[27] P. E. Gill, W. Murray, and M. A. Saunders, *SNOPT: An SQP algorithm for large-scale constrained optimization*, SIAM Rev., 47 (2005), pp. 99–131.
[28] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.
[29] N. I. M. Gould, M. E. Hribar, and J. Nocedal, *On the solution of equality constrained quadratic programming problems arising in optimization*, SIAM J. Sci. Comput., 23 (2001), pp. 1376–1395.
[30] N. I. M. Gould, S. Lucidi, M. Roma, and Ph. L. Toint, *Solving the trust-region subproblem using the Lanczos method*, SIAM J. Optim., 9 (1999), pp. 504–525.
[31] N. I. M. Gould, *On the accurate determination of search directions for simple differentiable penalty functions*, IMA J. Numer. Anal., 6 (1986), pp. 357–372.
[32] J. D. Griffin, *Interior-Point Methods for Large-Scale Nonconvex Optimization*, Ph.D. thesis, Department of Mathematics, University of California, San Diego, CA, 2005.
[33] W. W. Hager, *Minimizing a quadratic over a sphere*, SIAM J. Optim., 12 (2001), pp. 188–208.
[34] C. Keller, N. I. M. Gould, and A. J. Wathen, *Constraint preconditioning for indefinite linear systems*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1300–1317.
[35] L. Lukšan, C. Matonoha, and J. Vlček, *Interior-point method for non-linear non-convex optimization*, Numer. Linear Algebra Appl., 11 (2004), pp. 431–453.
[36] L. Lukšan and J. Vlček, *Indefinitely preconditioned inexact Newton method for large sparse equality constrained non-linear programming problems*, Numer. Linear Algebra Appl., 5 (1998), pp. 219–247.
[37] Y. Notay, *Flexible conjugate gradients*, SIAM J. Sci. Comput., 22 (2000), pp. 1444–1460.
[38] A. R. L. Oliveira and D. C. Sorensen, *A new class of preconditioners for large-scale linear systems from interior point methods for linear programming*, Linear Algebra Appl., 394 (2005), pp. 1–24.
[39] C. C. Paige and M. A. Saunders, *Solution of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629.
[40] I. Perugia and V. Simoncini, *Block-diagonal and indefinite symmetric preconditioners for mixed finite element formulations*, Numer. Linear Algebra Appl., 7 (2000), pp. 585–616.
[41] Y. Saad, *A flexible inner-outer preconditioned GMRES algorithm*, SIAM J. Sci. Comput., 14 (1993), pp. 461–469.
[42] M. A. Saunders and J. A. Tomlin, *Solving Regularized Linear Programs Using Barrier Methods and KKT Systems*, Report SOL 96-4, Department of EESOR, Stanford University, Stanford, CA, 1996.
[43] V. Simoncini and D. B. Szyld, *Flexible inner-outer Krylov subspace methods*, SIAM J. Numer. Anal., 40 (2003), pp. 2219–2239.
[44] T. Steihaug, *The conjugate gradient method and trust regions in large scale optimization*, SIAM J. Numer. Anal., 20 (1983), pp. 626–637.
[45] Ph. L. Toint, *Towards an efficient sparsity exploiting Newton method for minimization*, in Sparse Matrices and Their Uses, I. S. Duff, ed., Academic Press, London, New York, 1981, pp. 57–88.
[46] R. J. Vanderbei and T. J. Carpenter, *Symmetric indefinite systems for interior point methods*, Math. Program., 58 (1993), pp. 1–32.
[47] R. J. Vanderbei and D. F. Shanno, *An interior-point algorithm for nonconvex nonlinear programming*, Comput. Optim. Appl., 13 (1999), pp. 231–252.
[48] S. J. Wright and D. Orban, *Properties of the log-barrier function on degenerate nonlinear programs*, Math. Oper. Res., 27 (2002), pp. 585–613.
[49] S. J. Wright, *Primal-Dual Interior-Point Methods*, SIAM, Philadelphia, PA, 1997.

SIAM J. OPTIM.
Vol. 18, No. 2, pp. 691–709

DUALITY AND THE COMPUTATION OF APPROXIMATE INVARIANT DENSITIES FOR NONSINGULAR TRANSFORMATIONS*

CHRISTOPHER J. BOSE† AND RUA MURRAY‡

Abstract. We investigate a class of optimization problems which arise in the approximation of invariant densities for a nonsingular, measurable transformation T acting on a finite measure space. The problems under consideration have convex integral-type objectives and finite moment constraints and include, for example, the maximum entropy and quadratic programming approaches previously studied in the literature. This article is a natural sequel to those investigations and to the paper [C. Bose and R. Murray, *Discrete Contin. Dyn. Syst.*, 14 (2006), pp. 597–615], where a general class of convergent moment approximations were defined such that the limiting optimal solution is an invariant density for T. This article mainly concerns the solution of a single finite moment problem arising from this general approximation scheme. Both theoretical aspects and computational issues are treated. Although the problem fits easily into the standard theory of duality in convex optimization, its dynamical origins lead to technical obstructions in the derivation of optimality conditions. In particular, the dual functional for our problem is neither strictly convex nor coercive, relating in part to the fact that the moment generating functions for the approximation scheme need not be pseudo-Haar. The method of the paper circumvents these obstructions and yields an unexpected benefit: each finite moment approximation leads to rigorous bounds on the support of all invariant densities for T.

Key words. invariant measure, Frobenius–Perron operator, entropy-like objective, moment constraint, strong duality

AMS subject classifications. Primary, 28D05; Secondary, 37M25, 41A46, 49K27

DOI. 10.1137/060658163

1. Introduction.

Let $X = (X; \mathcal{B}, \mu)$ be a Borel measure space. When $T : X \to X$ is measurable and nonsingular with respect to μ, (T, X) is a dynamical system, and we are motivated by the question: *Can one find a T-invariant probability measure with a density function $f \in L^p(X; \mu)$ (usually $p = 1$)?* A measure ν is T-invariant, or an *invariant measure,* if $\nu = \nu \circ T^{-1}$. Invariant measures determine equilibrium statistics of the dynamical system (T, X) (via Birkhoff's ergodic theorem) and those with densities do so for a μ-nontrivial set of orbits. T-invariant measures with densities are *absolutely continuous invariant measures* (ACIMs). Usually, they cannot be found in closed form, and it is highly desirable to develop computational strategies for approximating them. In [5] we studied a class of convex optimization problems on classical Banach spaces whose solutions robustly approximate T-invariant densities: the solutions $\{f_n\}$ to appropriately chosen sequences of optimization problems (P_n) converge (in L^p) to a T-invariant density. In the current paper we investigate some of the technical issues that arise in solving such (P_n) as well as provide complete and explicit solutions for some special cases.

*Received by the editors April 25, 2006; accepted for publication (in revised form) February 23, 2007; published electronically August 24, 2007.
http://www.siam.org/journals/siopt/18-2/65816.html

†Dept. of Mathematics and Statistics, University of Victoria, P.O. Box 3045 STN CSC, Victoria, BC, Canada V8W 3P4 (cbose@math.uvic.ca). This author's research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

‡Dept. of Mathematics, University of Waikato, Private Bag 3015 Hamilton, New Zealand (r.murray@math.waikato.ac.nz).

The optimization problems have the general form

$$(P_n) \qquad \begin{aligned} &\text{Minimize } \Phi(f) = \int_X \phi(f(x))\, d\mu(x) \\ &\text{subject to } f \in L^p(X;\mu) \text{ and } Mf = \mathbf{b} \in \mathbb{R}^{N+1} \end{aligned}$$

(hereafter we denote $L^p(X;\mu) = L^p$). The constraint $M : L^p \to \mathbb{R}^{N+1}$ is of moment type, i.e., is defined with respect to a given finite collection[1] $g_0, g_1, \dots g_N$ of *moment test functions* in $L^q(X)$ and

$$(Mf)_i = \int f\, g_i\, d\mu, \quad i = 0, 1, \dots N.$$

For each n, the vector $\mathbf{b}$ is fixed and q is the conjugate index: $\frac{1}{p} + \frac{1}{q} = 1$ (when $p = 1$, $q = \infty$). This setup generalizes a study of Ding [6].

Our aim is to develop numerical algorithms for invariant density approximations which work in practice: the methods must produce a convergent sequence of approximately invariant densities, and each iteration must involve the computation of a well-defined function, which can be performed on a computer. Our optimization based program requires the following:

1. a suitable choice of *generating functions* (in L^q) such that any limit as $n \to \infty$ of solutions to (P_n) is an invariant density of the dynamical system (T, X); the dynamics of T are thus "encoded into M";
2. a choice of Φ which ensures norm convergence of the solutions of (P_n) as $n \to \infty$;
3. any refinements needed to ensure that the solution of (P_n) can be reduced to the solution of a finite number of algebraic equations; and
4. application of the method to specific examples to produce a convergent sequence of *approximately invariant densities*.

All of these steps lead to nontrivial considerations. Most of 1 is addressed in [5], where we refer the interested reader. The main requirement is that the moment test functions are derived from a sequence whose span is weak*-dense in L^q; the necessary details are given in section 2. The requirements of 2 can be addressed with standard results from the literature [1, 3, 12, 11], and some details are collected in section 2. Several example formulations are also presented in section 2.

The main effort in this paper is directed toward 3: establishing conditions which allow the *primal* optimization problems (P_n) to be solved on a computer. Since each (P_n) is convex, it is natural to write down the Lagrangian and pass to a *dual* (or conjugate) optimization problem, obtaining a concave, finite-dimensional, and unconstrained problem (D_n). While (D_n) is derived easily using standard methods, for a large (and reasonable) choice of moment formulations of the invariant measure problem, the dual objective function is noncoercive.[2] This leads to difficulty in the derivation of necessary and sufficient optimality conditions, and possibly to failure of dual attainment (with consequent impediments in the practical solution of the optimization problems). We have identified two mechanisms leading to noncoercivity of (D_n): (i) the moment test functions defining the constraint operators M may not be *pseudo-Haar*[3] leading to unbounded contours (in fact hyperplanes) in the

[1] Note that N may not equal n in some applications.
[2] That is, it has unbounded (upper) level sets; see [2].
[3] That is, they may not be linearly independent μ-almost everywhere.

objective of (D_n); and (ii) regions of X which are *transient* under the dynamics of T can prevent the dual problem (D_n) from attaining its maximum at all. Our main result (Theorem 3.3) is a condition which ensures dual attainment, leading to necessary and sufficient optimality conditions for the solution (Theorem 3.7). In the remainder of section 3 we develop a *domain restriction* which guarantees that the conditions in the theorems are met. Despite the ad hoc appearance of the restriction, it is intimately connected with the dynamics of T, and its imposition does not alter the solution of the underlying invariant measure problem. Moreover, the restriction yields useful dynamical information (see Lemma 3.8 (2)) which is not normally revealed by other methods for invariant density approximation (for example, Ulam's method [15, 9, 7, 10, 5]).

In section 4 we present several examples, illustrating how noncoercivity of the dual problems arises and how it is dealt with. In particular, we show how to accomplish the domain restriction for an "Entropy method" with simple moment test functions.

2. Optimization formulation of the invariant measure problem. The T-invariance condition for densities can be encoded into a sequence of constraint operators M for problems (P_n), and the optimization of Φ provides a convenient method of selecting a convergent sequence of approximately invariant measures. We now give a brief discussion of the dynamical origins of (P_n); further detail and discussion about connections between this and other methods for invariant measure approximation may be found in [5].

2.1. Encoding dynamics as moment constraints. Let (T, X) be a dynamical system. A σ-finite Borel measure ν is *absolutely continuous* if it has the form $d\nu = f d\mu$ for some measurable function[4] f. We write $\nu \ll \mu$ for absolute continuity and $f = \frac{d\nu}{d\mu}$. We also assume that T is *nonsingular*: $\mu \circ T^{-1} \ll \mu$, from which one can quickly deduce that $\nu \circ T^{-1} \ll \mu$ whenever $\nu \ll \mu$. As noted above, it is particularly interesting to find absolutely continuous probability measures which are invariant under T. For such a ν, $f = \frac{d\nu}{d\mu}$ is called an *invariant density*.

Invariant densities can be investigated via a transfer operator on L^p (usually, $p = 1$). If $d\nu = f d\mu$ then $\nu \circ T^{-1} \ll \mu$ so there is a function $\hat{f}$ satisfying $d(\nu \circ T^{-1}) = \hat{f} d\mu$. Thus, T induces an action P on L^1 by $Pf \stackrel{\text{def}}{=} \hat{f} = \frac{d(\nu \circ T^{-1})}{d\mu}$. The operator P is linear and positive (in the sense that $f \geq 0$ implies $Pf \geq 0$) and preserves integrals. P is called the *Frobenius–Perron* operator associated to T. Invariant densities $0 \leq f \in L^1$ are fixed points of P. An alternative[5] characterization of P is

$$\int Pf\, h\, d\mu = \int f\, h \circ T\, d\mu \quad \text{for all } f \in L^1,\ h \in L^\infty, \tag{2.1}$$

from which

$$Pf = f \text{ if and only if } \int f\,(h \circ T - h)\, d\mu = 0 \text{ for all } h \in L^\infty. \tag{2.2}$$

In view of (2.2), it is natural to express the invariant density condition via a sequence of moment approximations. Suppose $\mathcal{H} = \{h_1, h_2, \ldots h_N\} \subseteq L^\infty$ is a finite collection

[4]Normally we require $f \geq 0$, although signed absolutely continuous measures also make sense in this context. The usual definition of absolute continuity is $\mu(B) = 0 \Rightarrow \nu(B) = 0$; the equivalence of the two is part of the Lebesgue–Radon–Nikodym theorem.

[5]If B is a measurable set, then $\int Pf\, \mathbf{1}_B\, d\mu = \nu(T^{-1}B) = \int f\, \mathbf{1}_B \circ T\, d\mu$. For any simple h, linearity of the integral gives $\int Pf\, h\, d\mu = \int f\, h \circ T\, d\mu$, and the general case follows since simple functions are dense in L^∞.

of functions. We say that f is *approximately invariant up to* $\mathcal{H}$ if

$$\int f(h_i \circ T - h_i)\, d\mu = 0, \quad i = 1, 2, \dots N. \tag{2.3}$$

Setting $g_i = h_i \circ T - h_i$ and[6] $g_0 = \mathbf{1}$, we define the set of approximately invariant functions to be the *feasible set* for (P_n):

$$\mathcal{F}_n = \left\{ f \in L^1 \;\middle|\; \int f\, g_0\, d\mu = 1, \int f\, g_i\, d\mu = 0,\ i = 1, 2, \dots N \right\} = \left\{ f \in L^1 \mid Mf = \mathbf{b} \right\},$$

where $M : L^1 \to \mathbb{R}^{N+1}$ is defined by $(Mf)_i = \int f\, g_i\, d\mu$ and $\mathbf{b} = (1, 0, \dots, 0)$. (P_n) will be called *feasible* if $\mathcal{F}_n \neq \emptyset$, and each $f \in \mathcal{F}_n$ is *feasible for* (P_n). We call the collection $\mathcal{G} = \{g_0, g_1, g_2, \dots g_N\}$ the set of *moment test functions* and $\mathcal{H}$ the set of *generating functions* for the approximation. Notice that we do not explicitly include the non-negativity constraint $f \geq 0$ in the definition of the feasible set; we prefer to impose this, when desired, using the objective function Φ. (In [5] we present the case of $\Phi(f) = \frac{1}{2}\|f\|_{L^2}^2$ where allowing f to assume negative values is convenient.) The function g_0 ensures that the approximate invariant densities are normalized.[7] Note that each (P_n) has its own N, feasible set, moment test functions, and generating functions. When there is any possibility of ambiguity, these will be denoted $N(n), \mathcal{F}_n, \mathcal{G}_n$, and $\mathcal{H}_n$ (respectively). The possibility that $N \neq n$ should be emphasized. For example, when the moment generating functions arise from n-steps in a process of binary partition of the underlaying space, $N(n) = 2^n$. (See section 2.3 for this and other examples.)

Finally, we make the standing assumption that $\mu(X) < \infty$. Then, in the definition of $\mathcal{F}_n$, the function space L^1 can be replaced by $L^p,\ 1 < p < \infty$. This is due to the fact that $L^p \subseteq L^1,\ 1 \leq p < \infty$, and $L^\infty \subseteq L^q,\ 1 \leq q \leq \infty$, so all the integrals in $\mathcal{F}_n$ remain well defined. This allows us to consider a range of objectives Φ (such as H, V, and V^+ defined below).

Convergence of approximately invariant densities. The application of the problems (P_n) relies on the constraints being such that $\mathcal{F}_\infty = \bigcap_{n \geq 1} \mathcal{F}_n$ consists precisely of T-invariant densities. This condition is certainly satisfied when there exists an L^p T-invariant density, $\mathcal{H}_\infty = \{h_i\}_{i=1}^\infty$ is a sequence whose span is weak*-dense in L^q, and $\mathcal{H}_n = \{h_1, \dots, h_n\}$. In this situation, the sets $\{\mathcal{F}_n\}$ are *nested* ($\mathcal{F}_n \subseteq \mathcal{F}_m$ whenever $n \geq m$), so

$$\mathcal{F}_\infty = \left\{ f \;\middle|\; \int f\, d\mu = 1, \int (Pf - f)\, h\, d\mu = 0 \text{ for all } h \in \operatorname{span}(\mathcal{H}_\infty) \right\},$$

guaranteeing the condition in (2.2). (See [13] for generalizations of the L^q weak*-density condition.) The nested condition on $\mathcal{F}_n$ and the density condition on $\{\mathcal{H}_n\}$ can be weakened,[8] allowing other reasonable choices of $\{\mathcal{H}_n\}$. The role of the objective functional Φ is to specify a selection of $f_n \in \mathcal{F}_n$ such that $\lim_{n\to\infty} f_n$ exists and is in $\mathcal{F}_\infty$.

2.2. Choice of convex functional. The objective functionals Φ are chosen for mathematical and practical convenience. Let $\mu(X) < \infty$ and $\phi : X \to \mathbb{R} \cup \{\infty\}$ be *proper* [11], lower semicontinuous, and strictly convex. Define

$$\Phi(f) = \int \phi(f(x))\, d\mu(x).$$

[6] $\mathbf{1}_B$ denotes the characteristic function of the measurable subset B and $\mathbf{1} = \mathbf{1}_X$.

[7] This constraint also eliminates the trivial solution $f = 0$ from (P_n).

[8] In [5, section 3] we establish a suitable convergence result under "lattice" and "weak eventual clustering" conditions.

(So ϕ is a normal convex integrand in the sense of Rockafellar [11].) We require Φ to be strictly convex and weakly lower semicontinuous and to have weakly compact lower level sets and the *Kadec* property [3]: if $\Phi(f_n) \to \Phi(f) < \infty$ and $f_n \to f$ weakly as $n \to \infty$, then $\|f - f_n\|_{L^p} \to 0$.

Remarks.

1. The function $\phi(f)$ need not be integrable for every $f \in L^p$; however, we assume $[\phi(f)]^-$ (its negative part) is integrable, and consequently $\Phi(f)$ is unambiguously an element of $(-\infty, \infty]$. For the examples below, this assumption holds.
2. Provided there is an invariant density f_* for T with $f_* \in L^p$, (P_n) will be feasible for every choice of generators $\mathcal{H}$.

Natural choices for Φ. In [5] we studied the following choices for $\phi : \mathbb{R} \to \mathbb{R}$:

$$\phi(t) = \eta(t) \stackrel{\text{def}}{=} \begin{cases} t \log t & \text{for } t > 0, \\ 0 & \text{if } t = 0, \\ +\infty & \text{if } t < 0, \end{cases}$$

after which we adopt the standard notation $\Phi = H$, the (negative) Boltzmann–Shannon entropy on L^1. Notice that $H(f) < \infty$ implies that $f \geq 0$ μ-almost everywhere. If

$$\phi(t) = v(t) \stackrel{\text{def}}{=} \frac{1}{2} t^2,$$

we optimize with respect to an "Energy" functional which we denote by $\Phi = V$, and when

$$\phi(t) = v_+(t) \stackrel{\text{def}}{=} \begin{cases} \frac{1}{2} t^2 & \text{if } t \geq 0, \\ +\infty & \text{if } t < 0, \end{cases}$$

we get a positively constrained Energy functional, denoted $\Phi = V_+$. Of course the appropriate Banach space domains for the energy functionals V and V_+ would be L^2. Properties of these and other "Entropy-like" functionals are investigated in many papers (for example, [3, 2, 12]).

2.3. Three example setups.

Partition generating functions with "Energy" objective. Let $\mathcal{P}$ be a partition of X into measurable subsets $\mathcal{P} = \{B_i\}_{i=1}^N$, and let $\mathcal{H} = \{\mathbf{1}_{B_i}\}_{i=1}^N$ be the set of generating functions (in a slight abuse of terminology, we call this a *partition basis*). The moment test functions are therefore

$$g_i = \mathbf{1}_{B_i} \circ T - \mathbf{1}_{B_i} = \mathbf{1}_{T^{-1}B_i} - \mathbf{1}_{B_i},$$

and the approximately invariant densities can be considered as "invariant up to the discretization imposed by $\mathcal{P}$." Let $\Phi(f) = \frac{1}{2}\|f\|_{L^2}^2$. This example is studied in detail in [5] and leads to a convergent invariant density approximation scheme under the assumption that T admits an invariant density in L^2. The main complication that arises with this method is that $\sum_{i=1}^N g_i = \sum_{i=1}^N (h_i \circ T - h_i) = \mathbf{1}_X \circ T - \mathbf{1}_X = 0$ since $\sum_{i=1}^N h_i = \sum_{i=1}^N \mathbf{1}_{B_i} = \mathbf{1}_{\cup B_i} = \mathbf{1}_X$. Consequently, the M^* (defined below) has nontrivial kernel, leading to noncoercivity of the dual problem (D_n) (see section 3.2). This is dealt with easily, both analytically (section 3.2) and numerically [5,

Remark 7.1]. From a computational viewpoint, the case of a binary partition of the space is natural. Then we would have $N(n) = 2^n$ in the dimension of the moment problem n-steps into the associated approximation scheme.

Partition generating functions with "Entropy" objective. Again one uses a partition basis $\mathcal{H} = \{\mathbf{1}_{B_1}, \mathbf{1}_{B_2}, \dots \mathbf{1}_{B_N}\}$, but now $\Phi(f) = H(f) = \int_X \eta(f)\, d\mu = \int_X f(x) \log f(x)\, d\mu(x)$. For many interesting densities f (for example, the invariant densities for the logistic family of maps on $[0,1]$), $H(f) < \infty$ while $V(f) = \infty$. So, one gains applicability with this choice of Φ, but at cost: the dual optimization problem is potentially less tractable. In fact, the dual problem can suffer from noncoercivity, wherein the optimizer occurs "at infinity." In section 3.4 we elaborate and resolve these difficulties by restricting the domain of integration in (P_n).

Polynomial basis functions with "Entropy" objective. Let $X = [0,1] \subseteq \mathbb{R}$, let μ be Lebesgue measure, and let $\mathcal{H}_n = \{x, x^2, \dots x^n\}$. Here, quite naturally, $N(n) = n$. In Ding [6] this generating set is used, along with the entropy objective H to derive approximately invariant densities under the following dynamical assumptions:

(D1) The moment test functions $g_i(x) = (Tx)^i - x^i$, $i = 1, 2 \dots n$, are linearly independent; and

(D2) T admits a unique invariant density f_*, and further, this density satisfies $f_* > 0$ and $H(f_*) < \infty$.

In [4] we show, using techniques derived later in this article, that Ding's method can be extended to dynamical systems satisfying only the following:

(D3) T admits an invariant density f_* with $H(f_*) < \infty$ such that T is not of finite order with respect to $f_*\, d\mu$. (That is, there is no $n > 0$ so that $T^n = \mathrm{id}$, $f_*\, d\mu$-almost everywhere).

3. Main results.

3.1. The dual problem (D_n). Since ϕ is a *normal convex integrand* in the sense of Rockafellar [11], we have a simple closed form for the dual functional, which we denote by Q,

$$(D_n) \qquad \begin{aligned} &\text{Maximize } Q(\lambda) = \langle \lambda, \mathbf{b} \rangle - \int \phi^*([M^*\lambda](x))\, d\mu(x) \\ &\text{subject to } \lambda \in \mathbb{R}^{N+1}, \end{aligned}$$

where $M^* : \mathbb{R}^{N+1} \to L^q$ is the adjoint map defined by

$$M^*\lambda = \sum_i \lambda_i g_i \in L^q, \tag{3.1}$$

and where ϕ^* denotes the classical Fenchel (convex) conjugate of ϕ. Finally, weak duality holds:

$$\text{for all } \lambda \in \mathbb{R}^{N+1}, \text{ for all } f \in L^p \text{ such that } Mf = \mathbf{b},\ Q(\lambda) \le \Phi(f). \tag{3.2}$$

We refer readers not familiar with this type of argument to [2, 12] and provide a short, self-contained derivation of these facts in the appendix. The function ϕ^* is automatically convex (a fact we will need below), and the main work is in identifying conditions which guarantee that (D_n) attains its maximum at a finite λ (*dual attainment*); this is accomplished by proving that Q is coercive. As often occurs, dual attainment leads to necessary and sufficient conditions for both the dual and primal problems (section 3.3).

3.2. Dual attainment. A critical issue for solution of (P_n) is whether or not the dual problem (D_n) attains its maximum value. We remark at the outset that for many of our examples, the functional Q fails to be coercive. The treatment we give is motivated by [2], although immediate application of the results of that paper is impeded by the fact that our (P_n) do not necessarily admit feasible solutions in the quasirelative interior of L^p (the interesting T-invariant measures may not be supported on all of X). The first problem that can occur is that the operator M^* can have nontrivial kernel; this problem was noted in the partition basis examples above and is elaborated further in section 4.1 below.

LEMMA 3.1. *Suppose that (P_n) is feasible and write*[9] $\mathbb{R}^{N+1} = \mathrm{Ker}(M^*) \oplus \mathrm{Range}(M)$, *the canonical orthogonal direct sum. Then the following hold:*

1. $\mathbf{b} = (1, 0, 0, \ldots, 0) \in \mathrm{Range}(M)$.
2. $Q(\cdot)$ *is upper semicontinuous and constant on hyperplanes parallel to the subspace* $\mathrm{Ker}(M^*)$.
3. *If* $\mathrm{Ker}(M^*) \neq \{0\}$, *then Q is not coercive; however, in any event*
$$\mathrm{Max}_{\lambda \in \mathbb{R}^{N+1}} Q(\lambda) = \mathrm{Max}_{\lambda \in \mathrm{Range}(M)} Q(\lambda).$$
Moreover, (D_n) will attain its maximal value if and only if $Q|_{\mathrm{Range}(M)}$ attains its (relative) maximal value.
4. *Write* $\mathrm{Range}(M) = span\{\mathbf{b}\} \oplus \hat{\Lambda}$. *If f is feasible for (P_n) and $\hat{\lambda} \in \hat{\Lambda}$, then*
$$\int M^* \hat{\lambda}\, f \, d\mu = \langle \hat{\lambda}, Mf \rangle = \langle \hat{\lambda}, \mathbf{b} \rangle = 0$$
(where $\langle \cdot, \cdot \rangle$ is the usual inner product on $\mathbb{R}^{N+1}$).
5. *If $\lambda = \lambda_0 + \alpha \mathbf{b} + \hat{\lambda}$, where $\lambda_0 \in \mathrm{Ker}(M^*)$ and $\hat{\lambda} \in \hat{\Lambda}$, then*
$$Q(\lambda) = \alpha - \int \phi^*(\alpha\, \mathbf{1} + M^* \hat{\lambda})\, d\mu.$$

Proof. 1. Since (P_n) is feasible, there is an $f \in L^p$ for which $Mf = \mathbf{b}$. Thus $\mathbf{b} \in \mathrm{Range}(M)$. Statements 2–3 follow immediately from 1 and the formula for Q in (D_n). Statements 4–5 are direct computations. □

Given this lemma, dual attainment will follow once we have established that the upper level sets of $Q|_{\mathrm{Range}(M)}$ are bounded. This is done by exploiting the superlinear growth of the term $\int \phi^*(M^*(\cdot))\, d\mu$ (restricted to the linear subspace $\mathrm{Range}(M)$) to produce a bound on the decay of $Q(\lambda)$ as $\lambda \to \infty$.

LEMMA 3.2. *With notation as in Lemma* 3.1, *and with $\|\cdot\|$ denoting the Euclidean norm in $\mathbb{R}^{N+1}$, assume also the following:*

1. $\phi^* \geq 0$ *and* $\phi^*|_{[0,\infty)}$ *is nondecreasing.*
2. *For every $\hat{\lambda} \in \hat{\Lambda}$ with $\hat{\lambda} \neq 0$ one has $[M^* \hat{\lambda}]^+ \neq 0$.*

Then there exist $\gamma_0, \delta_0 > 0$ such that if $\lambda = \lambda_0 + \alpha \mathbf{b} + \hat{\lambda}$ and $\alpha + \|\hat{\lambda}\|\, \gamma_0 \geq 0$, then
$$Q(\lambda) \leq \alpha - \delta_0\, \phi^*(\alpha + \|\hat{\lambda}\|\, \gamma_0).$$

Proof. First, note that $\int [M^*(\cdot)]^+\, d\mu$ is continuous on $\mathbb{R}^{N+1}$ and, by hypothesis 2, is positive for every nonzero $\hat{\lambda} \in \hat{\Lambda}$. Since the unit sphere in $\hat{\Lambda}$ is compact in $\mathbb{R}^{N+1}$, there is a $\gamma > 0$ such that
$$\|\hat{\lambda}\| = 1 \;\Rightarrow\; \int_X [M^* \hat{\lambda}]^+\, d\mu \geq \gamma.$$

[9] Here, $\mathrm{Ker}(M^*) = \{\lambda \in \mathbb{R}^{N+1} \mid M^* \lambda = 0 \ \mu\text{-almost everywhere}\}$.

Let $0 < \gamma_0 < \frac{\gamma}{\mu(X)}$ and put $A_\lambda = \{x : [M^*(\frac{\hat{\lambda}}{\|\hat{\lambda}\|})](x) > \gamma_0\}$. Then

$$\gamma \le \int_X \left[M^* \tfrac{\hat{\lambda}}{\|\hat{\lambda}\|}\right]^+ d\mu = \int_{A_\lambda} \left[M^* \tfrac{\hat{\lambda}}{\|\hat{\lambda}\|}\right]^+ d\mu + \int_{X-A_\lambda} \left[M^* \tfrac{\hat{\lambda}}{\|\hat{\lambda}\|}\right]^+ d\mu$$
$$\le \|M^*\| \, [\mu(A_\lambda)]^{1/p} + [\mu(X)] \, \gamma_0,$$

where $\|M^*\|$ denotes the operator norm of $M^* : \mathbb{R}^{N+1} \to L^q$. We therefore conclude that

$$\mu(A_\lambda) \ge \left(\frac{\gamma - \mu(X)\,\gamma_0}{\|M^*\|}\right)^p \quad \stackrel{\text{def}}{=} \quad \delta_0. \tag{3.3}$$

Next, restricted to A_λ, $M^*(\lambda) = \alpha\,\mathbf{1} + M^*(\hat{\lambda}) \ge \alpha + \|\hat{\lambda}\|\,\gamma_0 \ge 0$. Thus

$$\frac{1}{\mu(A_\lambda)} \int_{A_\lambda} M^*\lambda \, d\mu \ge \alpha + \|\hat{\lambda}\|\,\gamma_0. \tag{3.4}$$

Now, since ϕ^* is convex, we have by Jensen's inequality

$$\mu(A_\lambda)\,\phi^*\left(\frac{1}{\mu(A_\lambda)} \int_{A_\lambda} M^*\lambda \, d\mu\right) \le \int_{A_\lambda} \phi^*(M^*\lambda)\, d\mu.$$

Since ϕ^* is nondecreasing, (3.3) and (3.4) lower bound the left-hand side by

$$\delta_0\,\phi^*(\alpha + \|\hat{\lambda}\|\,\gamma_0),$$

and since $\phi^* \ge 0$ we can upper bound the right-hand side to obtain

$$\delta_0\,\phi^*(\alpha + \|\hat{\lambda}\|\,\gamma_0) \le \int_X \phi^*(M^*\lambda)\, d\mu.$$

The lemma now follows from Lemma 3.1 (5). ∎

THEOREM 3.3. *With notation as in Lemma* 3.1, *assume that*

1. $\phi^* \ge 0$ *and* $\phi^*|_{[0,\infty)}$ *is nondecreasing;*
2. $\lim_{s\to+\infty} \frac{\phi^*(s)}{s} = \infty$*; and*
3. *for every* $\hat{\lambda} \in \hat{\Lambda}$ *with* $\hat{\lambda} \ne 0$ *one has* $[M^*\hat{\lambda}]^+ \ne 0$.

Then

$$\lim_{\|\lambda\|\to\infty,\ \lambda\in\mathrm{Range}(M)} Q(\lambda) = -\infty,$$

and the dual optimization problem (D_n) *attains its supremum.*

Proof. It suffices to establish that for any sequence $\{\lambda_n\} \subset \mathrm{Range}(M)$ with $\|\lambda_n\| \to \infty$, $Q(\lambda_n)$ is unbounded below. First, note that $\lambda_n = \alpha_n\,\mathbf{b} + \hat{\lambda}_n$. If any subsequence $\{\lambda_{n_i}\}$ has $\alpha_{n_i} \to -\infty$, then $Q(\lambda_{n_i}) \le \alpha_{n_i} \to -\infty$ by Lemma 3.1 (5) (recall that $\phi^* \ge 0$). Thus, we need only consider sequences $\{\lambda_n\}$ for which $\{\alpha_n\}$ is bounded below. If $\{\alpha_n\}$ is also bounded above, then since $\|\lambda_n\| \to \infty$, we must have $\lim_{n\to\infty} \|\hat{\lambda}_n\| \to \infty$ so that $\alpha_n + \|\hat{\lambda}_n\|\,\gamma_0 \to \infty$. In particular, $\alpha_n + \|\hat{\lambda}_n\|\,\gamma_0 \ge 0$ for all large enough n, so by Lemma 3.2,

$$Q(\lambda_n) \le \alpha_n - \delta_0\,\phi^*(\alpha_n + \|\hat{\lambda}_n\|\,\gamma_0) \to -\infty$$

(since $\lim_{s\to\infty}\phi^*(s)=\infty$). The only other possibility is that $\{\alpha_n\}$ is unbounded above, in which case there is a subsequence $\{\lambda_{n_j}\}$ of $\{\lambda_n\}$ for which $\lim_{j\to\infty}\alpha_{n_j}=\infty$. Then, in view of hypothesis 2, there is an N such that

$$\alpha_{n_j}\geq 0 \quad\text{and}\quad \frac{\phi^*(\alpha_{n_j})}{\alpha_{n_j}}\geq\frac{2}{\delta_0} \quad\text{for } j\geq N.$$

Then use Lemma 3.2 to estimate

$$Q(\lambda_{n_j})\leq\alpha_{n_j}-\delta_0\,\phi_*(\alpha_{n_j})\leq-\alpha_{n_j}\to-\infty \quad\text{as}\quad j\to\infty. \qquad \square$$

Note. The limit in condition 2 could be replaced by a lim sup since ϕ^* is convex. Theorem 3.3 is analogous to [2, Theorem 4.8], but condition 3 is unnecessary there due to the assumption of a strictly positive feasible point for (P_n).

Example 3.4. Suppose $X=[0,1]$, μ is Lebesgue measure, and

$$T(x)=\begin{cases}2x & \text{if } 0\leq x<1/2,\\ 2(x-\frac{1}{2})+\frac{1}{2} & \text{if } \frac{1}{2}\leq x<\frac{3}{4},\\ 2(x-\frac{1}{2}) & \text{if } \frac{3}{4}\leq x\leq 1.\end{cases}$$

Then $f_*=2\mathbf{1}_{[\frac{1}{2},1]}$ is the unique invariant probability density for T. Let $\mathcal{P}=\{[0,\frac{1}{2}),[\frac{1}{2},1]\}$. The functions g_i are

$$g_0=\mathbf{1},\qquad g_1=-\mathbf{1}_{[\frac{1}{4},\frac{1}{2})},\qquad g_2=\mathbf{1}_{[\frac{1}{4},\frac{1}{2})},$$

so $\mathrm{Ker}(M^*)=\mathrm{span}\{(0,1,1)^T\}$, $\mathrm{Range}(M)=\mathrm{span}\{(1,0,0)^T,(0,1,-1)^T\}$, and $\hat{\Lambda}=\mathrm{span}\{(0,1,-1)^T\}$. But evidently, $M^*(0,1,-1)^T=-2\mathbf{1}_{[\frac{1}{4},\frac{1}{2})}\leq 0$, so that hypothesis 3 of Theorem 3.3 fails. In fact, using the entropy functional H as the objective, one easily computes the dual functional Q on the two-dimensional subspace $\mathrm{Range}(M)$ (where vectors take the form $(\alpha,\beta,-\beta)^T$) as

$$Q((\alpha,\beta,-\beta)^T)=\alpha-e^{(\alpha-1)}\left\{\frac{1}{4}e^{-2\beta}+\frac{3}{4}\right\}.$$

So Q is noncoercive, $\sup_{\mathrm{Range}(M)}Q=\log(4/3)$, but it is not reached at any point of $\mathrm{Range}(M)$, so dual attainment fails.

Example 3.5. We can immediately apply Theorem 3.3 to establish coercivity of Q for Ding's polynomial basis maximum entropy method [6]. The condition (D1) (above) implies that $\mathrm{Ker}(M^*)=\{0\}$, so the decomposition in Lemma 3.1 is $\mathbb{R}^{n+1}=\mathrm{Range}(M)$ and the set $\hat{\Lambda}=\{\lambda\in\mathbb{R}^{n+1}\mid\lambda_0=0\}=(\mathrm{span}\{\mathbf{b}\})^\perp$. Now suppose $\lambda\neq 0$ and $\lambda^T\mathbf{b}=0$ so that $M^*(\lambda)\neq 0$. If $[M^*\lambda]^+=0$ then $M^*\lambda=[M^*\lambda]^-$ (almost everywhere), so whenever $f>0$ is feasible for (P_n),

$$\langle\lambda,\mathbf{b}\rangle=\langle\lambda,Mf\rangle=\int M^*\lambda f\,d\mu<0.$$

Since (D2) guarantees the existence of a feasible, almost everywhere positive invariant density f_*, this calculation contradicts $\lambda^T\mathbf{b}=0$. We conclude that $[M^*\lambda]^+\neq 0$ and Theorem 3.3 yields dual attainment. Even without condition (D1), the restriction of (D_n) to $\mathrm{Range}(M)$ will yield dual attainment by the same argument.

3.3. Necessary and sufficient optimality conditions. Once dual attainment is established, the dual (D_n) and primal (P_n) problems are linked by a standard derivation of optimality conditions [2, 11]. We begin by quoting a calculus lemma.

LEMMA 3.6. *Assume $\mu(X) < \infty$ and that $\varphi : \mathbb{R} \to \mathbb{R}$ with $\varphi \in C^1$. Suppose $A : \mathbb{R}^{N+1} \to L^\infty(X)$ is linear and set, for every $\mathbf{z} \in \mathbb{R}^{N+1}$,*

$$q(\mathbf{z}) = \int_X \varphi((A\mathbf{z})(x))\, d\mu(x).$$

Then

1. *for every $\mathbf{z}_0 \in \mathbb{R}^{N+1}$, $\varphi'(A\mathbf{z}_0(\cdot)) \in L^1$;*
2. *for every $\mathbf{z}_0 \in \mathbb{R}^{N+1}$, $(\nabla_{\mathbf{z}} A)\mathbf{z}_0(\cdot) \in [L^\infty]^{N+1}$; and*
3. *q is Gateaux differentiable at every $\mathbf{z}_0 \in \mathbb{R}^{N+1}$ and in particular*

$$(\nabla_{\mathbf{z}} q(\mathbf{z}_0))_i = \int \varphi'(A\mathbf{z}_0(x)) Ae_i(x)\, d\mu(x) \in \mathbb{R}.$$

THEOREM 3.7 (necessary and sufficient optimality conditions). *Assume that the primal problem (P_n) is feasible and $\varphi = \phi^*$ is smooth and satisfies the hypothesis of Lemma 3.6. If $\lambda(n)$ yields a global maximum of Q in the dual formulation (D_n), then*

1. *$\lambda(n)$ satisfies*

$$\int [\phi^*]'([M^*\lambda(n)](x)) g_i(x)\, d\mu(x) = \mathbf{b}_i, \quad i = 0, 1, 2 \ldots n; \tag{3.5}$$

2. *$f_n = [\phi^*]'(M^*\lambda(n)) \in L^1$ is feasible for (P_n); and*
3. *$Q(\lambda(n)) = \int \phi(f_n(x))\, d\mu(x)$, and hence, from the weak duality condition (3.2), we conclude that f_n is a minimizer in the primal problem (P_n).*

In particular, (3.5) is also sufficient for an optimal value of λ in (D_n) and the function f_n defined in part 2 is optimal in the primal problem.

Proof. Using Lemma 3.6 we establish necessary conditions for $\lambda(n)$ to maximize Q:

$$0 = -\mathbf{b}_i + \int [\phi^*]'([M^*\lambda(n)](x)) g_i(x)\, d\mu(x), \qquad i = 0, 1, 2 \ldots N,$$

so f_n defined in part 2 satisfies $f_n \in L^1$ and the constraint $Mf_n = \mathbf{b}$. Now, since ϕ^* is convex, proper, and smooth, one easily derives from classical facts (see the appendix) that for all $s \in \mathbb{R}$

$$\phi^*(s) + \phi^{**}([\phi^*]'(s)) = s[\phi^*]'(s),$$

which, combined with $\phi^{**} = \phi$, yields

$$\phi([\phi^*]'(s)) + \phi^*(s) = s[\phi^*]'(s).$$

If we now substitute $s = [M^*\lambda(n)](x)$ and rearrange to obtain

$$\phi(f_n(x)) = [M^*\lambda(n)](x) f_n(x) - \phi^*([M^*\lambda(n)](x)),$$

we see that $\phi(f_n(\cdot))$ is an integrable function since both functions on the right are integrable. Conclude that f_n is feasible for (P_n). Finally, integrating this last expression over $x \in X$ yields

$$\Phi(f_n) = Q(\lambda(n)),$$

closing the duality gap and proving both that f_n is a minimizer of Φ in (P_n) and that $\lambda(n)$ is a maximizer of Q in (D_n) if and only if (3.5) holds. ∎

3.4. Domain restriction with a partition basis. In Example 3.4, the existence of a nonpositive, nonzero $M^*\lambda$ prevented Q from being coercive and destroyed any prospect of dual attainment. However, $\text{supp}(M^*\lambda)$ was contained in a part of X that was transient under the action of T. In general, if f is feasible[10] for (P_n) and $M^*\lambda \leq 0$, then $\text{supp}(M^*\lambda) \cap \text{supp}(f) = \emptyset$, so any noncoercivity of Q because of failure of condition 3 in Theorem 3.3 can be attributed to the behavior of M^* on an unimportant part of X. Motivated by this (and justified in Lemma 3.8 and section 4.2 below), we employ a (P_n)-specific *domain restriction*.

Consider the (sub)cone of $\mathbb{R}^{N+1}$ defined by $\mathcal{C} = \{\hat{\lambda} \in \hat{\Lambda} \mid M^*\hat{\lambda} \leq 0\}$ and set $X_0(n) = X \setminus \bigcup_{\hat{\lambda}\in\mathcal{C}}\{x \in X \mid [M^*\hat{\lambda}](x) < 0\}$.

LEMMA 3.8. *Let $\mathcal{P}$ be a finite measurable partition of X and let the constraints in (P_n) be with respect to the corresponding partition basis. Then the following hold:*

1. *$X_0(n)$ belongs to the σ-algebra generated by $\mathcal{P} \vee T^{-1}\mathcal{P}$ (and so is measurable).*
2. *Assume that ϕ satisfies the hypothesis of Theorem 3.3 and that $\Phi(f) < \infty \implies f \geq 0$ almost everywhere. Then $\Phi(f) < \infty$, and f feasible for (P_n) implies $\text{supp}(f) \subseteq X_0(n)$.*
3. *Under the same condition as in part 2, $X_0(n)$ is not a null-set of X and the value of the problem*

$$(P_n') \qquad \begin{aligned} &\textit{Minimize} \quad \Phi_0(f) = \int_{X_0(n)} \phi(f(x))\, d\mu(x) \\ &\textit{subject to} \quad f \in L^p(X_0(n)) \quad \textit{and} \quad Mf = \mathbf{b} \in \mathbb{R}^{N+1} \end{aligned}$$

is identical to the value of (P_n). Dual attainment holds in the case of the problem (P_n'),

$$Q_0(\lambda) \stackrel{def}{=} \text{Max}_{\lambda\in\mathbb{R}^{N+1}}\{\langle\lambda, \mathbf{b}\rangle - \textstyle\int_{X_0(n)} \phi^*(M^*(\lambda))\, d\mu\}.$$

Proof. 1. Observe that for each λ, $\{x | M^*\lambda(x) < 0\}$ is an element of the *finite* σ-algebra generated by $\mathcal{P} \vee T^{-1}\mathcal{P}$ (cf. Lemma 4.1 below). It follows that $X_0(n)$ is measurable, even though the union is over an uncountable parameter set. 2. When $\Phi(f) < \infty$ and f is feasible for (P_n), $\int M^*\hat{\lambda} f\, d\mu = 0$ for all $\hat{\lambda}$ by Lemma 3.1 (4). This implies that $\text{supp}(f) \subseteq X_0(n)$. 3. Since $f_0 \in L^p(X_0(n))$ is feasible for (P_n') if and only if $f_0\mathbf{1}_{X_0(n)} \in L^p(X)$ is feasible for (P_n), either both problems are infeasible, or there is a feasible $f \neq 0$. In this case, $\int_{X_0(n)} f\, d\mu = 1$, so $X_0(n) \neq \emptyset$. Furthermore, $\Phi_0(f) = \Phi(f)$ for all feasible f. Dual attainment holds since restricted to $X_0(n)$, hypothesis 3 of Theorem 3.3 holds. □

In effect, we have moved troublesome vectors λ where $M^*\lambda \leq 0$ into $\text{Ker}(M^*)$ over the restricted measure space $X_0(n)$. Of course, the domain for (P_n) is therefore changed, as is Φ, but our argument shows that the values of the two problems are identical, and the restricted problem has dual attainment.

Example 3.4 *revisited.* Recall that dual attainment failed due to the noncoercivity of Q. However, observe that if $0 \neq \hat{\lambda} \in \hat{\Lambda}$ and $M^*\hat{\lambda} \leq 0$, then we have $\text{supp}(M^*\hat{\lambda}) = [\frac{1}{4}, \frac{1}{2})$ (note that $M^*\hat{\lambda} \neq 0$ since $\hat{\lambda} \in \text{Range}(M) = (\text{Ker}(M^*))^\perp$). By Lemma 3.1 (4) $\int M^*\hat{\lambda} f\, d\mu = 0$ and $f \geq 0$ (provided f is feasible). Thus, $\text{supp}(f) \subseteq ([0,1] \setminus [1/4, 1/2))$, so we let $X_0 = ([0,1] \setminus [1/4, 1/2))$ and solve

$$\text{Minimize} \quad H_0(f) = \int_{X_0} \eta(f(x))\, d\mu(x) \quad \text{s.t.} \quad f \in L^1(X_0) \text{ and } Mf = \mathbf{b}.$$

[10] For example, any T-invariant density.

4. Applications. We now discuss some of the computational details when the method is applied to the specific case of a partition basis. For the rest of the paper, we assume for simplicity that $N(n) = n$ and let $\mathcal{P} = \{B_1, \ldots, B_n\}$ be the measurable partition defining the basis.

4.1. The "Energy" method with kernel. Recall that $\phi(t) = \frac{1}{2}t^2$. Then, since

$$Q(\lambda) = \langle \lambda, \mathbf{b} \rangle - \frac{1}{2} \int (M^*\lambda(x))^2 \, d\mu(x)$$

and the conjugate in the second term is weakly (in fact, norm) lower semicontinuous, Q is norm upper semicontinuous, so it attains its supremum over compact subsets. Let

$$\mathbb{R}^{n+1} = \mathbb{R}^{N+1} = \mathrm{Ker}(M^*) \oplus \mathrm{Range}(M),$$

the canonical decomposition relative to the operator M. This decomposition is nontrivial since $\sum_{i=1}^n g_i = \sum_{i=1}^n \mathbf{1}_{T^{-1}B_i} - \mathbf{1}_{B_i} = \mathbf{1} - \mathbf{1} = 0$, so $(0, 1, 1, \ldots 1)^T \in \mathrm{Ker}(M^*)$ (cf. Lemma 3.1.) Clearly, if Q restricted to the subspace $\mathrm{Range}(M)$ attains its (relative) maximum at λ^*, then Q will also be maximized at λ^*. To see why dual attainment holds in this case, note that

$$\begin{aligned} \mathrm{Max}_\lambda Q(\lambda) = \mathrm{Max}_{\lambda \in \mathrm{Range}(M)} Q(\lambda) &= \mathrm{Max}_{\lambda \in \mathrm{Range}(M)} \left\{ \langle \lambda, \mathbf{b} \rangle - \frac{1}{2} \textstyle\int (M^*\lambda)^2 \, d\mu(x) \right\} \\ &= \mathrm{Max}_{\lambda \in \mathrm{Range}(M)} \left\{ \langle \lambda, \mathbf{b} \rangle - \frac{1}{2} \langle \lambda, MM^*\lambda \rangle \right\}. \end{aligned}$$

The linear operator MM^* maps $\mathrm{Range}(M)$ into $\mathrm{Range}(M)$ and for $\lambda \neq 0$ in $\mathrm{Range}(M)$ we have $\langle \lambda, MM^*\lambda \rangle > 0$ so the operator $MM^*|_{\mathrm{Range}(M)}$ is positive definite. It follows that the restricted functional $Q|_{\mathrm{Range}(M)}$ is a negative definite quadratic form, is therefore coercive, and attains its maximum value.

There is no need to identify the restricted measure space $X_0(n)$ from this point of view, and applying Theorem 3.7 yields the necessary equation for the optimal value of $\lambda(n)$,

$$\sum_j [\lambda(n)]_j \textstyle\int g_i(x) \, g_j(x) \, d\mu(x) = b_i, \quad i = 0, 1, \ldots n, \tag{4.1}$$

and the formula for the optimal solution f_n,

$$f_n = M^*\lambda(n) = \sum_j [\lambda(n)]_j g_j. \tag{4.2}$$

Since $[\phi^*]'(s) = s$, the equation to be solved is linear and consistent in $n + 1$ variables:

$$A[\lambda(n)] = \mathbf{b},$$

where $A = \{a_{ij}\}$ is the $(n+1)\times(n+1)$ matrix of correlations: $a_{ij} = \int g_i(x) g_j(x) \, d\mu(x)$. Notice that $A = MM^*$ with $\mathrm{Ker}(A) = \mathrm{Ker}(MM^*) = \mathrm{Ker}(M^*)$, along which we know Q is constant, so *any* solution of (4.1) will lead to optimal values for both primal and dual (see also Theorem 3.7). In section 4.3 we present results of some numerical experiments concerning this problem with respect to the basis $\phi_i = \mathbf{1}_{B_i}$ generated by a partition.

Further details (including some issues about numerical implementation) are in [5].

4.2. The "Entropy" method and domain restriction. In the case of a partition basis, the X_0 of Lemma 3.8 may be needed to ensure dual attainment. We show below how to identify $X_0(n)$ by a finite computation. Once this is done, Theorem 3.7 can be invoked to derive the optimality equations in concrete form:

$$\int_{X_0(n)} \exp\{[M^*\lambda(n)](x) - 1\} g_i(x)\, d\mu(x) = \mathbf{b}_i, \quad i = 0, 1, \dots n, \tag{4.3}$$

from which the primal optimal points will be computed according to the formula in Theorem 3.7(2). That is, we recover the solution to (P_n) by solving (P_n') with $\lambda(n)$ satisfying (4.3). The solution to (P_n) is then $f_0(x) = \mathbf{1}_{X_0(n)} \exp\{[M^*\lambda(n)](x) - 1\}$.

Identification of restricted domain. We now let $\mathcal{P} = \{B_1, \dots, B_n\}$ be a fixed partition of X. Thus, $n = N$ is fixed, and we suppress, where possible, explicit notational dependence on n. For example, from now on we will write X_0 instead of $X_0(n)$. Recall the decomposition $\text{Range}(M) = \{\mathbf{b}\} \oplus \hat{\Lambda}$. Then $\mathcal{C} = \{\lambda \in \hat{\Lambda} | M^*\lambda \le 0\}$ and $X_0 = X \setminus \bigcup_{\lambda \in \mathcal{C}} \{x | M^*\lambda(x) < 0\}$.

LEMMA 4.1. *For each $i, j = 1, \dots, n$,*

$$M^*\lambda|_{B_i \cap T^{-1}B_j} = (\lambda_0 + \lambda_j - \lambda_i)\mathbf{1}_{B_i \cap T^{-1}B_j}.$$

Proof. Since both $\{B_k\}_{k=1}^n$ and $\{T^{-1}B_k\}_{k=1}^n$ are partitions of X, the lemma follows directly from the facts that $M^*\lambda = \sum_{k=0}^n \lambda_k g_k$ and $\mathbf{1}_{B_i \cap T^{-1}B_j} = \mathbf{1}_{B_i}\mathbf{1}_{T^{-1}B_j}$. ∎

LEMMA 4.2. *Let A be the $(n \times n)$ matrix with entries $A_{ij} = \mu(B_i \cap T^{-1}B_j)$, and let $\lambda \in \hat{\Lambda}$ be such that $M^*\lambda \le 0$. If $(A^{m_1})_{ij} > 0$ and $(A^{m_2})_{ji} > 0$ for some $m_1, m_2 > 0$, then $\lambda_i = \lambda_j$.*

Proof. Since $\lambda \in \hat{\Lambda}$, $\lambda_0 = 0$. Since each $A_{kl} \ge 0$, there is a sequence $\{i_k\}_{k=0}^{m_1+m_2}$ such that $i_0 = i = i_{m_1+m_2}$, $i_{m_1} = j$ and each $A_{i_l i_{l+1}} > 0$. Then, by Lemma 4.1,

$$(\lambda_{i_{l+1}} - \lambda_{i_l})\, A_{i_l i_{l+1}} = \int (\lambda_{i_{l+1}} - \lambda_{i_l}) \mathbf{1}_{B_{i_l} \cap T^{-1}B_{i_{l+1}}}\, d\mu = \int_{B_{i_l} \cap T^{-1}B_{i_{l+1}}} M^*\lambda\, d\mu \le 0.$$

Thus $\lambda_{i_0} \ge \lambda_{i_1} \ge \cdots \ge \lambda_{i_{m_1}} \ge \cdots \ge \lambda_{i_{m_1+m_2}} = \lambda_{i_0}$. In particular, $\lambda_i = \lambda_{i_0} = \lambda_{i_{m_1}} = \lambda_j$. ∎

PROPOSITION 4.3. *The following are equivalent:*

(i) *$A_{ij} > 0$ and $(A^m)_{ji} > 0$ for some $m > 0$;*

(ii) *$\mu(B_i \cap T^{-1}B_j) > 0$ and $B_i \cap T^{-1}B_j \subset X_0$ (mod μ).*

Proof. (i)⇒(ii) Suppose that $\mu(B_i \cap T^{-1}B_j) = A_{ij} > 0$, $(A^m)_{ji} > 0$, and let $\lambda \in \mathcal{C}$. Then, using Lemma 4.2 with $m_1 = 1$ and $m_2 = m$ gives $\lambda_i = \lambda_j$. By Lemma 4.1, $M^*\lambda(x) = \lambda_j - \lambda_i = 0$ when $x \in B_i \cap T^{-1}B_j$. This establishes that $B_i \cap T^{-1}B_j \subset X_0$.

(ii)⇒(i) We assume that $\mu(B_i \cap T^{-1}B_j) = A_{ij} > 0$ but that $(A^m)_{ji} = 0$ for all $m > 0$. We need to construct a $\hat{\lambda} \in \mathcal{C}$ such that $M^*\hat{\lambda}|_{B_i \cap T^{-1}B_j} < 0$, since this will show that $B_i \cap T^{-1}B_j$ is disjoint from X_0 μ-almost everywhere. Let $\mathcal{I} = \{j\} \cup \{k : (A^m)_{jk} > 0 \text{ for some } m > 0\}$ and define λ by putting $\lambda_l = -\mathbf{1}_{\mathcal{I}}(l)$ and $\lambda_0 = 0$. Observe that (a) $\lambda_i = 0$ and $\lambda_j = -1$; and (b) if $k \in \mathcal{I}$ and $A_{kl} > 0$ then $l \in \mathcal{I}$. Now, by Lemma 4.1, if $A_{kl} > 0$ then $M^*\lambda|_{B_k \cap T^{-1}B_l} = \lambda_l - \lambda_k$. By observation (a), $M^*\lambda|_{B_i \cap T^{-1}B_j} = -1$. We now check that $M^*\lambda \le 0$: by observation (b), if $\lambda_k = -1$ and $A_{kl} > 0$ then $\lambda_l = -1$ so $M^*\lambda|_{B_k \cap T^{-1}B_l} = 0$; on the other hand, if $\lambda_k = 0$ then $\lambda_l - \lambda_k \le 0$, so in any event $M^*\lambda \le 0$. Finally, decompose $\lambda = \hat{\lambda} + z$, where $\hat{\lambda} \in \hat{\Lambda}$ and $z \in \text{Ker}(M^*)$. Then $M^*\hat{\lambda} = M^*\lambda \le 0$ and $M^*\hat{\lambda}|_{B_i \cap T^{-1}B_j} < 0$. ∎

Proposition 4.3 suggests an elementary iterative procedure for identifying X_0 up to a set of measure 0:

1. Calculate the $n \times n$ matrix $A_{ij} = \mu(B_i \cap T^{-1}B_j)$.
2. For each $A_{ij} > 0$, determine $\mathcal{I}(j) = \{k | (A^m)_{jk} > 0 \text{ for some } m > 0\}$. If $i \in \mathcal{I}(j)$, then $B_i \cap T^{-1}B_j \subset X_0$, and set $\hat{A}_{ij} := A_{ij}$. Otherwise, set $\hat{A}_{ij} := 0$.

At the end of this procedure, set $\mathcal{I}_0 = \{(i,j) | \hat{A}_{ij} > 0\}$. Then take

$$X_0 = \cup_{(i,j)\in\mathcal{I}_0} B_i \cap T^{-1}B_j.$$

Remarks 4.4.

1. For reasonably regular maps T the matrix A is very sparse, with $O(n)$ nonzero entries which can be stored as a list of triples (i, j, A_{ij}). Consequently each set $\mathcal{I}(j)$ can be determined in $O(n)$ operations (mostly array look-ups); the identification of $\mathcal{I}_0$ via the above procedure thus requires at most $O(n^2)$ operations.
2. Proposition 4.3 essentially characterizes X_0 as elements of the partition $\mathcal{P} \vee T^{-1}\mathcal{P}$ which correspond to *strongly connected components* of a certain directed graph.[11] If A has $O(n)$ nonzero entries, all of these components (and the edges connecting them) can be found with $O(n)$ computational effort by Tarjan's algorithm [14]. See [8] for related work on the use of discrete models to obtain recurrent components and Lyapunov functions of dynamical systems.

The following corollary to Proposition 4.3 will be used below.

COROLLARY 4.5. *If* $(\hat{A}^m)_{ik} > 0$ *then there is an* $M > 0$ *such that* $(\hat{A}^M)_{ki} > 0$.

Proof. There are indices $i = i_0, i_1, \ldots, i_m = k$, and integers $M_1, \ldots, M_m$ such that $\hat{A}_{i_{l-1}i_l} > 0$ and $(\hat{A}^{M_l})_{i_l i_{l-1}} > 0$ for $l = 1, \ldots, m$. Then

$$(\hat{A}^{M_1+\cdots+M_m})_{ki} \geq (\hat{A}^{M_m})_{i_m i_{m-1}} \cdots (\hat{A}^{M_1})_{i_1 i_0} > 0. \qquad \square$$

Solution of the necessary conditions. Since the solution to (P_n) is obtained via (D_n), one needs to maximize

$$Q(\lambda) = \langle \lambda, \mathbf{b} \rangle - \int_{X_0} \exp\{M^*\lambda(x) - 1\}\, d\mu(x).$$

Using Lemma 4.1 and Proposition 4.3, we have

$$Q(\lambda) = \lambda_0 - \exp\{\lambda_0 - 1\} \sum_{(i,j)\in\mathcal{I}_0} \hat{A}_{ij} \exp\{\lambda_j - \lambda_i\},$$

so that (D_n) is solved by minimizing $G(\lambda) = \sum_{(i,j)\in\mathcal{I}_0} \hat{A}_{ij} \exp\{\lambda_j - \lambda_i\}$ and setting $\lambda_0 = 1 - \log(\sum_{(k,l)\in\mathcal{I}_0} \hat{A}_{kl} \exp\{\lambda_l - \lambda_k\})$. The optimal values of λ can then be used to recover the solution to (P_n) as in Theorem 3.7(2). The minimum of $G(\lambda)$ can be calculated using standard optimization algorithms, although we obtained rapid convergence with a fixed point method that we now describe.

The equations $\frac{\partial G}{\partial \lambda_i} = 0$ reduce to $\sum_l \hat{A}_{il} \exp\{\lambda_l - \lambda_i\} = \sum_k \hat{A}_{ki} \exp\{\lambda_i - \lambda_k\}$. Thus, for $i = 1, \ldots, n$,

$$(e^{-\lambda_i})^2 = \frac{\sum_{k\neq i} \hat{A}_{ki}\, e^{-\lambda_k}}{\sum_{l\neq i} \hat{A}_{il}\, e^{\lambda_l}},$$

[11] The vertices are the elements of $\mathcal{P}$ and the edge set corresponds to those ij with $A_{ij} > 0$.

which suggests an iterative scheme $(\lambda_i)^{(m+1)} = -\frac{1}{2}\log F_i(e^{-\lambda_1^{(m)}}, \ldots, e^{-\lambda_n^{(m)}})$ with the choice $F_i(x_1, \cdots, x_n) = \frac{\sum_{k\neq i}\hat{A}_{ki}\,x_k}{\sum_{l\neq i}\hat{A}_{il}\,\frac{1}{x_l}}$. In practice, it is more convenient to work directly with the values $x_i^{(m)} = e^{-(\lambda_i)^{(m)}}$, updating according to

$$x_i^{(m+1)} = \frac{\sqrt{F_i(\mathbf{x}^{(m)})}}{\sum_j \sqrt{F_j(\mathbf{x}^{(m)})}}.$$

We have no general proof for convergence of this iteration but note that it worked in all cases we tested, using $(x_i)^{(0)} = 1$.

Remarks 4.6.

1. The definition of F_i needs slight modification to allow for the possibilities that (i) $\sum_{l\neq i}\hat{A}_{il} = 0$ or (ii) $x_l = 0$. In the case of (i), Corollary 4.5 ensures that $\sum_{k\neq i}\hat{A}_{ki} = 0$, from which it follows that $G(\lambda)$ is independent of λ_i. In this case, set $F_i(\mathbf{x}) := 1$. In the case of (ii), an indeterminate expression is obtained only for those i with $\hat{A}_{il} > 0$, and continuity of F_i can then be ensured by putting $F_i(\mathbf{x}) = 0$.
2. The normalization of $\mathbf{x}^{(m+1)}$ ensures that the iteration scheme preserves the unit simplex in $(\mathbb{R}_+)^n$, without altering the value of $G(\lambda)$ (if $\mathbf{x} \mapsto c\,\mathbf{x}$, the effect on $x_i = e^{-\lambda_i}$ is $\lambda_i \mapsto \lambda_i - \log c$, and for any $\log c \in \mathbb{R}$, $G(\lambda) = G(\lambda - \log c)$).
3. The iteration is not a uniform contraction of the unit simplex since it preserves the boundary.

4.3. Numerical examples. We now apply the energy and entropy minimization approaches to approximate the invariant measures for several examples on the unit interval. μ in this case is Lebesgue measure.

Example 1. Let

$$T(x) = \begin{cases} 2\,x, & x \in [0, 1/2), \\ 2x - 1/2, & x \in [1/2, 3/4), \\ 2\,x - 1, & x \in [3/4, 1]. \end{cases}$$

The invariant measure for T has density $f_*(x) = 2\,\mathbf{1}_{[1/2,1]}$. For a sequence of values of n, approximations $f_n^{(V)}, f_n^{(H)}$ have been calculated which minimize $V(f) = \frac{1}{2}\int f^2\,d\mu$ and $H(f) = \int f\log f\,d\mu$, respectively. The approximation errors $\|f - f_n^{(V)}\|_{L^1}$ and $\|f - f_n^{(H)}\|_{L^1}$ are displayed in Table 1. The density approximations for $n = 729$ are displayed in the first row of Figure 1. Notice that $f_{729}^{(V)}$ has some negative values in $[0, 1/2)$; this is possible because our formulation of the optimization problem (with V) imposes no positivity condition although $[f_n]_- \to 0$ and $[f_n]_+ \to f$ (see [5, Remark 5.3(2)]). The spikes in $f_{729}^{(H)}$ occur at preimages of $\frac{1}{2}$ (a boundary point of $\mathrm{supp}(f_*)$) and disappear when $\frac{1}{2}$ is a boundary of a B_i.

Example 2. Let

$$T(x) = \begin{cases} 3\,x, & x \in [0, 1/4), \\ x + 1/2, & x \in [1/4, 1/2), \\ x - 1/2, & x \in [1/2, 3/4), \\ 3\,x - 2, & x \in [3/4, 1]. \end{cases}$$

TABLE 1

L^1 approximation errors for energy and entropy minimization approaches to invariant density calculations.

	Example 1		Example 2	
n	$\|f - f_n^{(V)}\|_{L^1}$	$\|f - f_n^{(H)}\|_{L^1}$	$\|f - f_n^{(V)}\|_{L^1}$	$\|f - f_n^{(H)}\|_{L^1}$
3	0.666667	0.699359	0.098485	0.104804
9	0.346007	0.326124	0.063600	0.067046
27	0.196225	0.134116	0.042583	0.043930
81	0.090671	0.051596	0.035391	0.036633
243	0.038662	0.019901	0.027982	0.029287
729	0.021628	0.006858	0.024727	0.025740
2187	0.008310	0.002605	0.022186	0.023187
6561	0.003775	0.000863	0.020258	0.021252

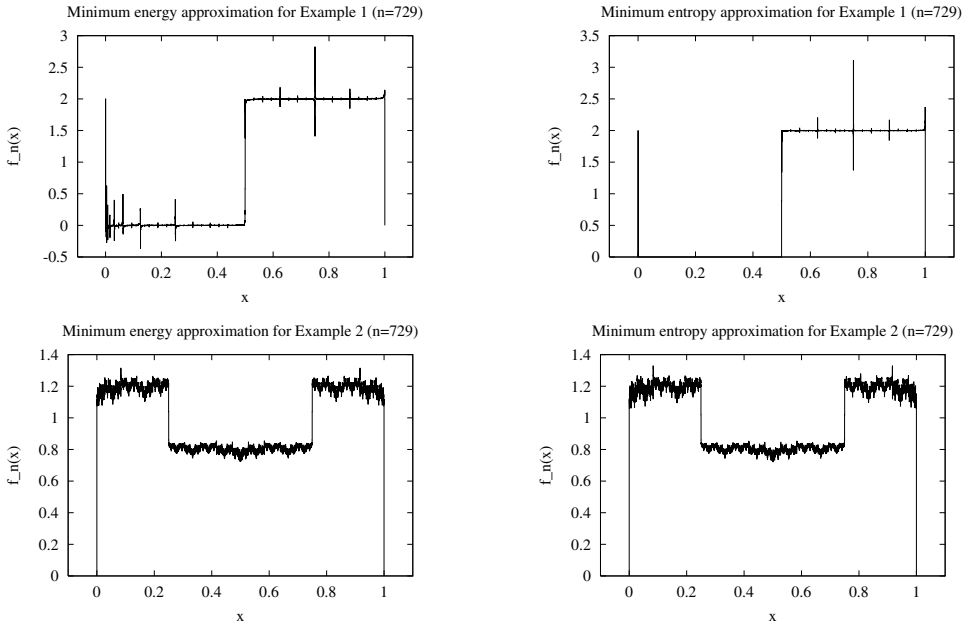


FIG. 1. *$n = 729$ density approximations for the maps in Examples 1 and 2 using Energy and Entropy minimization.*

The invariant measure for T has density $f_*(x) = 1.2\,\mathbf{1}_{[0,1/4)\cup(3/4,1]} + 0.8\mathbf{1}_{[1/4,3/4]}$. For a sequence of values of n, approximations $f_n^{(V)}, f_n^{(H)}$ have been calculated which minimize $V(f) = \frac{1}{2}\int f^2\,d\mu$ and $H(f) = \int f \log f\,d\mu$, respectively. The approximation errors $\|f - f_n^{(V)}\|_{L^1}$ and $\|f - f_n^{(H)}\|_{L^1}$ are displayed in Table 1. The density approximations for $n = 729$ are displayed in the second row of Figure 1.

Example 3. The tent map $T_r(x) = r(0.5 - |x - 0.5|)$ admits an invariant density f_r (of bounded variation) whenever $r \in (1, 2]$. Therefore, $H(f_r) < \infty$, and a sequence of f_n solving the finitely constrained optimization problems (P_n) will converge in L^1 to f_r as $n \to \infty$. In fact, if $r \in (2^{2^{-(k+1)}}, 2^{2^{-k}})$, then the density is supported on a union of 2^k intervals. In Figure 2, the $n = 729$ minimum entropy approximation is displayed for the tent map with $r = 1.3$. The displayed density is supported on $X_0(n)$, a union

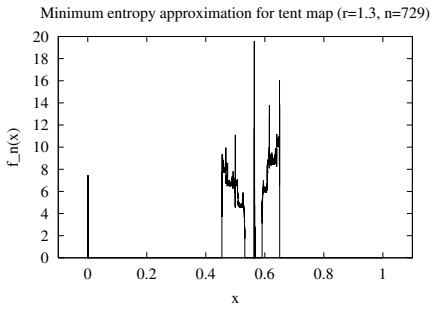


FIG. 2. *Entropy minimization with* $n = 729$ *for density approximation for the tent map (Example* 3*).*

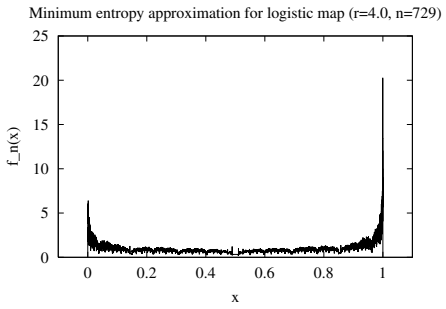


FIG. 3. *Entropy minimization with* $n = 729$ *for density approximation for the fully developed logistic map (Example* 4*).*

of several intervals, the larger two contain the support of the invariant density for T_r, and the remaining (small) intervals are clustered near the unstable fixed points for T_r at $x = 0, \frac{r}{1+r} \approx 0.565$. The correct density has no simple formula at this parameter value, so a direct calculation of the approximation error is not possible.

Example 4. The logistic map $T_r(x) = r\,x\,(1-x)$ admits an invariant density $f_*(x) = \frac{1}{\pi\sqrt{x\,(1-x)}}$ when $r = 4.0$. Then, $H(f_*) = 0.241564\cdots < \infty$, so the minimum entropy method will produce a sequence of density approximations f_n such that $\lim_{n\to\infty} \|f_n - f\|_{L^1} = 0$, even though neither T_r, nor any of its iterates, is expanding. The $n = 729$ minimum entropy approximation is displayed in Figure 3. (The error $\|f_n - f_*\|_{L^1} = 0.24683$.)

Appendix. Derivation of (D_n). The Lagrangian for (P_n) is

$$L(f,\lambda) = \Phi(f) - \langle \lambda, Mf - \mathbf{b}\rangle, \quad f \in L^p,\ \lambda \in \mathbb{R}^{N+1},$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product in $\mathbb{R}^{N+1}$. Next, define

$$\begin{aligned} Q(\lambda) &= \inf_{f\in L^p} L(f,\lambda) \\ &= \langle \lambda, \mathbf{b}\rangle - \sup_{f\in L^p}\{\langle \lambda, Mf\rangle - \Phi(f)\} \\ &= \langle \lambda, \mathbf{b}\rangle - \Phi^*(M^*\lambda), \end{aligned}$$

where $\Phi^* : L^q \to \mathbb{R}$ denotes the Fenchel (convex) conjugate of Φ, that is,

$$\Phi^*(g) = \sup_{f \in L^p} \left\{ \int f(x)g(x)\, d\mu(x) - \Phi(f) \right\},$$

and the adjoint $M^* : \mathbb{R}^{N+1} \to L^q$ is calculated as

$$M^*\lambda = \sum_{k=0}^{n} \lambda_k g_k.$$

We note that Φ^* is easily seen to be both convex and weakly lower semicontinuous on L^q.

The functional Φ^* is the Banach space generalization of the classical convex conjugate ϕ^* for real functions $\phi : \mathbb{R} \to (-\infty, \infty]$. See Rockafellar [12] for definitions and elementary properties. When Φ is of integral type, there are some important connections between the two concepts.

For example, if $f \in L^p$ is such that $\phi(f(\cdot))$ is integrable, then for all $g \in L^q$, from Fenchel's inequality $\phi(t) + \phi^*(s) \geq ts$, after substituting $t = f(x)$ and $s = g(x)$ and integrating, one obtains $\int \phi^*(g(x))\, d\mu(x) \geq \int f(x)g(x)\, d\mu(x) - \int \phi(f(x))\, d\mu(x)$. It follows that $\int \phi^*(g(x))\, d\mu(x) \in (-\infty, \infty]$ unambiguously and $\int \phi^*(g(x))\, d\mu(x) \geq \Phi^*(g)$. These and many other properties of integral-type objectives are derived in Rockafellar [11]. We summarize the facts that we will use.

LEMMA A.1. *Let $\phi : \mathbb{R} \to (-\infty, \infty]$ be a convex, lower semicontinuous, and proper function.*

1. *Suppose that for every $f \in L^p$, $\Phi(f) = \int \phi(f(x))\, d\mu(x)$ is unambiguously an element of $(-\infty, \infty]$. Then for each $g \in L^q$, $\int \phi^*(g(x))\, d\mu(x)$ is unambiguously defined as an element of $(-\infty, \infty]$ and*

$$\Phi^*(g) = \int \phi^*(g(x))\, d\mu(x).$$

2. *If $\phi^*(g)$ is integrable for at least one $g \in L^q$ then the integral $\int \phi(f(x))\, d\mu(x)$ is well defined (possibly $= \infty$) for every $f \in L^p$. This will be the case, for example, if ϕ^* is proper and $\mu(X) < \infty$.*

Equipped with these tools, we can write down the dual optimization problem associated to (P_n) as

$$(D_n) \qquad \begin{aligned} &\text{Max} \quad Q(\lambda) = \langle \lambda, \mathbf{b} \rangle - \int_X \phi^*(M^*\lambda)(x)\, d\mu(x) \\ &\text{subject to} \quad \lambda \in \mathbb{R}^{N+1}, \end{aligned}$$

an unconstrained, finite-dimensional, and concave problem. It follows directly from the definitions of Q and L that

$$Q(\lambda) \leq \inf_{f \in L^p,\, Mf = \mathbf{b}} L(f, \lambda) \leq \Phi(f) \tag{A.1}$$

whenever f is feasible for (P_n) and $\lambda \in \mathbb{R}^{N+1}$. Hence, the (maximal) value of (D_n) is majorized by the (minimal) value of (P_n), the so-called principle of *weak duality*. Thus, solving the unconstrained dual problem (D_n) is equivalent to solving the primal (P_n) precisely when this "duality gap" can be closed. Theorem 3.7 describes one

situation which is tailored to our applications and where the duality gap can be closed.

Acknowledgments. Both authors would like to thank the Department of Mathematics and Statistics, University of Victoria, and the Department of Mathematics, University of Waikato, for hospitality during the period when this research was conducted.

REFERENCES

[1] J. M. Borwein and A. S. Lewis, *Convergence of best entropy estimates*, SIAM J. Optim., 1 (1991), pp. 191–205.
[2] J. M. Borwein and A. S. Lewis, *Duality relationships for entropy-like minimization problems*, SIAM J. Control. Optim., 29 (1991), pp. 325–338.
[3] J. M. Borwein and A. S. Lewis, *On the convergence of moment problems*, Trans. Amer. Math. Soc., 325 (1991), pp. 249–271.
[4] C. Bose and R. Murray, *Dynamical conditions for convergence of a maximum entropy method for Frobenius–Perron operator equations*, Appl. Math. Comput., 182 (2006), pp. 210–212.
[5] C. Bose and R. Murray, *Minimum 'energy' approximations of invariant measures for nonsingular transformations*, Discrete Contin. Dyn. Syst., 14 (2006), pp. 597–615.
[6] J. Ding, *A maximum entropy method for solving Frobenius–Perron operator equations*, Appl. Math. Comput., 93 (1998), pp. 155–168.
[7] J. Ding and A. Zhou, *Finite approximations of Frobenius–Perron operators. A solution of Ulam's conjecture to multi–dimensional transformations*, Phys. D, 92 (1996), pp. 61–68.
[8] W. D. Kalies, K. Mischaikow, and R. C. A. M. VanderVorst, *An algorithmic approach to chain recurrence*, Found. Comput. Math., 5 (2005), pp. 409–449.
[9] T.-Y. Li, *Finite approximation for the Perron–Frobenius operator. A solution to Ulam's conjecture*, J. Approx. Theory, 17 (1976), pp. 177–186.
[10] R. Murray, *Approximation error for invariant density calculations*, Discrete Contin. Dyn. Syst., 4 (1998), pp. 535–558.
[11] R. T. Rockafellar, *Integrals which are convex functionals*, Pacific J. Math., 24 (1968), pp. 525–539.
[12] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[13] H. H. Schaefer, *Topological vector spaces*, Grad. Texts in Math. 3, Springer-Verlag, New York, 1970.
[14] R. Tarjan, *Depth-first search and linear graph algorithms*, SIAM J. Comput., 1 (1972), pp. 146–160.
[15] S. Ulam, *A Collection of Mathematical Problems*, Interscience Publishers, New York, 1960.

SPECIAL ISSUE ON VARIATIONAL ANALYSIS AND OPTIMIZATION

This issue of *SIAM Journal on Optimization* was motivated by the September 2005 Workshop on Well-Posedness of Optimization Problems and Related Topics, held in Borovets, Bulgaria. This workshop was the tenth event of a series initiated in 1987 as a small meeting of Bulgarian and Italian mathematicians working on the subject. Since then, this series has gained a high international reputation, and its workshops are attractive scientific events where modern research of leading experts is presented and lively discussions stimulate and motivate young scientists. The workshops have substantially increased their scope by including stability, sensitivity, and well-posedness of problems in optimization, optimal control, and calculus of variations, variational principles, advances in stochastic optimization, vector optimization, set-valued analysis, and so on.

We view variational analysis as a large area of modern mathematics encompassing not only ideas of the classical calculus of variations and optimization but also to a large extent perturbations and approximations, set-valued analysis, generalized differential calculus, and generalized convexity. Interaction with other branches of mathematics and with engineering and economics brought interesting and unexpected applications of variational analysis and nonsmooth optimization. Recent developments in mathematical finance, actuarial mathematics, and statistics have been major driving forces for the theory and numerical methods of optimization.

This volume presents 23 papers authored or co-authored by the participants. The topics reflect the field's diversity and vitality. They highlight many interesting recent advances and indicate some problems that may become focus of future research.

Many people have made this collection possible. We would like to express our gratitude to all organizations who supported the Borovets workshop and to the local organizers who created the productive and supportive atmosphere of the meeting. We are grateful to the authors for their contributions and to the referees for their excellent and timely work. Finally, we offer our thanks to the SIAM Vice-President for Publications, Tim Kelley; to the Editor-in-Chief of *SIAM Journal on Optimization*, Nick Gould; and SIAM staffers Mitch Chernoff and Brian Fauth, who have encouraged us to prepare this special issue and have supported us in the process of editing it.

We hope that this collection of papers provides an interesting mathematical lecture and an inspiration for further research.

Darinka Dentcheva
Stevens Institute of Technology
Hoboken, NJ

Julian Revalski
Institute of Mathematics and Informatics, Bulgarian Academy of Sciences
Sofia, Bulgaria
guest editors

SIAM J. OPTIM.
Vol. 18, No. 3, pp. 711–716

UNIQUENESS AND COMPARISON RESULTS FOR FUNCTIONALS DEPENDING ON ∇u AND ON u^*

A. CELLINA†

Abstract. We introduce a family of solutions to a variational problem and we prove uniqueness and comparison results.

Key words. strict convexity, comparison theorem, uniqueness

AMS subject classification. 49K20

DOI. 10.1137/060657455

1. Introduction.

We consider the problem of minimizing the functional

$$\mathcal{J}(u) = \int_\Omega [f(\nabla u(x)) + \alpha u(x)]\, dx, \tag{1}$$

where $\alpha \neq 0$ is a constant, on a suitable class of functions. We shall say that a function $w \in W^{1,1}(\Omega)$ is a *solution* if the functional attains its minimum among all the functions in $W^{1,1}(\Omega)$ that satisfy the same boundary conditions as w. Our purpose is to provide a comparison result and a result on uniqueness of solutions. A comparison result is a statement of the following kind: "For w and v solutions (satisfying different boundary conditions), $w \leq v$ on $\partial\Omega$ implies $w \leq v$ on Ω." A uniqueness result instead concerns solutions with the same boundary datum.

The assumptions we make on the convex function f are very general and apply to convex Lagrangeans that can be extended valued and not necessarily differentiable; moreover, f can be such that the domain of the subdifferential is strictly smaller than the domain of the function. Apart from this, our main point is that we do not assume that f is strictly convex: in this sense, this paper is a sequel to [1]. Notice, however, that the uniqueness result we present does not hold no matter what the boundary data are (a result of this kind would be rather unlikely, without any assumption of strict convexity), but holds only for a restricted class of boundary conditions.

The boundary conditions are those satisfied by a family of solutions that we are going to describe. We emphasize the fact that this family is explicit, i.e., that it can be computed directly from the Lagrangean f.

2. A family of solutions.

In what follows, f^* is the *polar* of f; for its main properties, we refer to [3] and [2].

THEOREM 1. *Let Ω be an open bounded set, regular enough so that the divergence theorem holds, and let $f : R^N \to R \cup \{+\infty\}$ be an extended valued, convex, lower semicontinuous function. For x_0 and θ in R^N and $c \in R$, consider the function*

$$\omega_\alpha(x) = \frac{N}{\alpha} f^* \left(\theta + \frac{x - x_0}{N} \alpha \right) + c.$$

*Received by the editors April 18, 2006; accepted for publication (in revised form) October 26, 2006; published electronically October 4, 2007.
http://www.siam.org/journals/siopt/18-3/65745.html
†Dipartimento di Matematica e Applicazioni, Universitá degli Studi di Milano-Bicocca, Via R. Cozzi 53, 20125 Milano, Italy (arrigo.cellina@unimib.it).

If ω_α is defined on Ω and belongs to $W^{1,1}(\Omega)$, then it is the only minimum of the functional

$$\mathcal{J}(u) = \int_\Omega [f(\nabla u(x)) + \alpha u(x)]\, dx,$$

in the class of functions

$$\mathcal{S} = \left\{ u \in W^{1,1}(\Omega), u - \omega_\alpha \in W_0^{1,1}(\Omega) \right\}.$$

As an example, the polar to the convex function

$$f(\xi) = \begin{cases} \|\xi\| - \sqrt{\|\xi\|} & \text{if } \|\xi\| \geq \sqrt{1}, \\ 0 & \text{if } \|\xi\| \leq \sqrt{1} \end{cases}$$

is given by

$$f^*(p) = \begin{cases} \frac{3}{4(1-\|p\|)} & \text{if } \frac{1}{2} \leq \|p\| < 1, \\ \|p\| & \text{if } \|p\| \leq \frac{1}{2}. \end{cases}$$

For $N = 1$, $\alpha = 1$, $\Omega = (-1, 1)$, $\theta = 0$, $x_0 = 0$, the function

$$\omega(x) = f^*(x)$$

is defined on Ω but does not belong to $W^{1,1}(\Omega)$. However, one has the following result.

Remark 1. If, in addition, f has superlinear growth, then, for Ω bounded, for every x_0 and θ, the function $\omega_\alpha(x)$ is defined on $\overline{\Omega}$ and belongs to $W^{1,\infty}(\Omega)$.

Indeed, by the superlinear growth of f, we obtain that $f^*(p) < +\infty$ for every $p \in R^N$. In particular, for every θ and x_0 in R^N, the convex function $\omega_\alpha(x)$ is defined and Lipschitzian on $\overline{\Omega}$.

The proof of Theorem 1 is direct and does not depend on the validity of the Euler–Lagrange equation, since, to this author's knowledge, the validity of this equation, for the class of problems considered in this paper, is yet to be established. We shall need the following lemma.

LEMMA 1. *Let f be convex, let $\xi_1 \neq \xi_2$, and let p be in both $\partial f(\xi_1)$ and $\partial f(\xi_2)$. Then, f^* is not differentiable at p.*

Proof. Since $p \in \partial f(\xi_2)$, for $\lambda > 0$ we have

$$\frac{1}{\lambda}(f^*(p + \lambda(\xi_2 - \xi_1)) - f^*(p))$$

$$= \frac{1}{\lambda}\left(\sup_\xi \{\langle p + \lambda(\xi_2 - \xi_1), \xi\rangle - f(\xi)\} - (\langle p, \xi_2\rangle - f(\xi_2)) \right) \geq \langle (\xi_2 - \xi_1), \xi_2\rangle.$$

Analogously, since $p \in \partial f(\xi_1)$, we have

$$\frac{1}{\lambda}(f^*(p + \lambda(\xi_1 - \xi_2)) - f^*(p)) \geq \langle (\xi_1 - \xi_2), \xi_1\rangle$$

so that, if f^* is differentiable at p, we obtain

$$\langle (\xi_2 - \xi_1), \xi_1\rangle \geq \langle \nabla f^*(p), \xi_2 - \xi_1\rangle \geq \langle (\xi_2 - \xi_1), \xi_2\rangle,$$

i.e., $\|\xi_2 - \xi_1\|^2 \leq 0$. □

Proof of Theorem 1. Consider the case $\alpha > 0$; in this case, ω_α is a convex function (it would be concave for $\alpha < 0$).

(a) We shall prove that for every $u \in \mathcal{S}$, we have $\mathcal{J}(\omega_\alpha) \leq \mathcal{J}(u)$. By assumption, the effective domain of the convex function ω_α contains Ω; hence, ω_α is locally Lipschitzian, hence differentiable almost everywhere, on Ω. Let $x \in \Omega$ be such that $\nabla\omega_\alpha(x)$ exists, and set $p(x) = \theta + \frac{x-x_0}{N}\alpha$. Then $f^*(p)$ is differentiable in p at $p = p(x)$, with gradient $z = \nabla f^*(p) = \nabla f^*(p(x))$. Since $z \in \nabla f^*(p)$ implies $p \in \partial f(z)$, we obtain that

$$p(x) \in \partial f\left(\nabla f^*\left(p(x)\right)\right) = \partial f(z),$$

so that, for every $q \in R^N$,

$$f(q) - f(z) \geq \langle p(x), q - z\rangle .$$

We have that

$$\nabla\omega_\alpha(x) = \nabla f^*\left(\theta + \frac{x-x_0}{N}\alpha\right) = \nabla f^*(p(x)) = z.$$

Hence, in particular, for $u(x) \in \mathcal{S}$ and for almost every $x \in \Omega$, we obtain

$$f\left(\nabla u(x)\right) - f\left(\nabla\omega_\alpha(x)\right) \geq \langle p(x), \nabla u(x) - \nabla\omega_\alpha(x)\rangle , \tag{2}$$

so that

$$\int_\Omega \left[f\left(\nabla u(x)\right) - f\left(\nabla\omega_\alpha(x)\right)\right] dx \geq \int_\Omega \left\langle \theta + \frac{\alpha}{N}(x-x_0), \nabla u(x) - \nabla\omega_\alpha(x)\right\rangle dx.$$

We have that $u - \omega \in W_0^{1,1}(\Omega)$; recalling the divergence theorem, we obtain

$$\int_\Omega \langle \theta, \nabla u(x) - \nabla\omega_\alpha(x)\rangle \, dx = \int_\Omega \operatorname{div}((u(x) - \omega_\alpha(x))\theta)\, dx = 0$$

and

$$\begin{aligned}
&\int_\Omega \left\langle \frac{\alpha}{N}(x-x_0), \nabla u(x) - \nabla\omega_\alpha(x)\right\rangle dx \\
&= \int_\Omega \left[\operatorname{div}\left((u(x) - \omega_\alpha(x))\frac{\alpha}{N}(x-x_0)\right) - \frac{\alpha}{N}(u(x) - \omega_\alpha(x))\operatorname{div}(x-x_0)\right] dx \\
&= -\int_\Omega \alpha(u(x) - \omega_\alpha(x))\, dx,
\end{aligned}$$

so that

$$\int_\Omega \left[f\left(\nabla u(x)\right) - f\left(\nabla\omega_\alpha(x)\right)\right] dx \geq -\int_\Omega \alpha\left(u(x) - \omega_\alpha(x)\right) dx,$$

i.e.,

$$\int_\Omega \left[f\left(\nabla u(x)\right) + \alpha u(x)\right] dx \geq \int_\Omega \left[f\left(\nabla\omega_\alpha(x)\right) + \alpha\omega_\alpha(x)\right] dx.$$

Since u is arbitrary in $\mathcal{S}$, the above inequality shows that ω_α is a solution.

(b) Let w be another solution. Because it is a solution, we have

$$\int_\Omega \left([f(\nabla w(x)) - f(\nabla \omega_\alpha(x))] + \alpha[w(x) - \omega_\alpha(x)]\right) dx = 0. \tag{3}$$

Since

$$-\int_\Omega \langle p(x), \nabla w(x) - \nabla \omega_\alpha(x)\rangle \, dx = -\int_\Omega \left\langle \theta + \frac{x - x_0}{N}\alpha, \nabla w(x) - \nabla \omega_\alpha(x) \right\rangle dx$$

$$= \alpha \int_\Omega [w(x) - \omega_\alpha(x)] \, dx,$$

from (3) we obtain

$$\int_\Omega \left([f(\nabla w(x)) - f(\nabla \omega_\alpha(x))] - \langle p(x), \nabla w(x) - \nabla \omega_\alpha(x)\rangle\right) dx = 0.$$

From (2) applied to the solution w, we have that

$$[f(\nabla w(x)) - f(\nabla \omega_\alpha(x))] - \langle p(x), \nabla w(x) - \nabla \omega_\alpha(x)\rangle \geq 0,$$

so that, for a.e. $x \in \Omega$, we obtain

$$f(\nabla w(x)) - f(\nabla \omega_\alpha(x)) - \langle p(x), \nabla w(x) - \nabla \omega_\alpha(x)\rangle = 0.$$

Next, we claim that $p(x) \in \partial f(\nabla w(x))$ as well. In fact, for every ξ, from the above equality we obtain

$$f(\xi) - f(\nabla w(x)) = f(\xi) - [f(\nabla \omega_\alpha(x)) + \langle p(x), \nabla w(x) - \nabla \omega_\alpha(x)\rangle]$$

$$= f(\xi) - [f(\nabla \omega_\alpha(x)) + \langle p(x), \xi - \nabla \omega_\alpha(x)\rangle + \langle p(x), \nabla w(x) - \xi\rangle] \geq \langle p(x), \nabla w(x) - \xi\rangle$$

so that, by definition, $p(x) \in \partial f(\nabla w(x))$.

Apply Lemma 1. By assumption, we have that f^* is differentiable at $p = p(x)$; hence we infer that $\nabla w(x) = \nabla \omega_\alpha(x)$. The point x was arbitrary and the two functions satisfy the same boundary condition so that $w = \omega_\alpha$. □

The following is our comparison result; here, by saying that at $\partial\Omega$ we have $v \geq u$, we mean, as usual, that $(u - v)^+ \in W_0^{1,1}(\Omega)$.

Corollary 1 (comparison theorem). *Let w be a solution to the minimization of* (1) *in the class of those functions satisfying the same boundary conditions as w; assume that, on $\partial\Omega$, we have $w \leq \omega_\alpha$. Then, $w \leq \omega_\alpha$ a.e. in Ω.*

Proof. Set $\eta = (w - \omega_\alpha)^+$ so that $\eta \in W_0^{1,1}(\Omega)$. We have that $\tilde{\omega}_\alpha = \omega_\alpha + \eta$ is such that $\tilde{\omega}_\alpha - \omega_\alpha \in W_0^{1,1}(\Omega)$, while, defining $\tilde{w} = w - \eta$, we have that $\tilde{w}$ satisfies $\tilde{w} - w \in W_0^{1,1}(\Omega)$. Set $E^+ = \{x \in \Omega : \eta(x) > 0\}$: on E^+, $\nabla \tilde{\omega}_\alpha = \nabla w$ and $\tilde{\omega}_\alpha = w$, while $\nabla \tilde{w} = \nabla \omega_\alpha$ and $\tilde{w} = \omega_\alpha$.

Since ω_α is a solution on the set $\omega - \omega_\alpha \in W_0^{1,1}(\Omega)$, we have

$$0 \leq \int_\Omega [f(\nabla \tilde{\omega}_\alpha(x)) + \alpha \tilde{\omega}_\alpha(x)] \, dx - \int_\Omega [f(\nabla \omega_\alpha(x)) + \alpha \omega_\alpha(x)] \, dx$$

$$= \int_{E^+} [f(\nabla w(x)) + \alpha w(x)] \, dx - \int_{E^+} [f(\nabla \omega_\alpha(x)) + \alpha \omega_\alpha(x)] \, dx.$$

In the same way, since w is a solution on the set $v - w \in W_0^{1,1}(\Omega)$, we obtain

$$0 \leq \int_{E^+} [f(\nabla \omega_\alpha(x)) + \alpha \omega_\alpha(x)]\, dx - \int_{E^+} [f(\nabla w(x)) + \alpha w(x)]\, dx$$

so that

$$\int_{E^+} [f(\nabla \omega_\alpha(x)) + \alpha \omega_\alpha(x)]\, dx = \int_{E^+} [f(\nabla w(x)) + \alpha w(x)]\, dx$$
$$= \int_{E^+} [f(\nabla \tilde{\omega}_\alpha(x)) + \alpha \tilde{\omega}_\alpha(x)]\, dx$$

and $\tilde{\omega}_\alpha$ is a further solution to the minimization of (1) on $\{u : u - \omega_\alpha \in W_0^{1,1}(\Omega)\}$. This solution differs from ω_α when $m(E^+) > 0$, a contradiction to Theorem 1. Hence, $m(E^+) = 0$. □

3. The limit as $\alpha \to 0$. The functions ω_α are undefined for $\alpha = 0$. We are interested in the question of whether the functions ω_α converge to a limit as $\alpha \to 0$, and, if this is the case, of how this limit is related to the solutions of the minimum problem with $\alpha = 0$, i.e., of the problem

$$\text{minimize} \int_\Omega f(\nabla u(x))\, dx, \qquad u - u_0 \in W_0^{1,1}(\Omega). \tag{4}$$

In particular, write the arbitrary constant c as $c = -\frac{N}{\alpha} f^*(\theta) + \beta$ (β arbitrary) and consider the family of solutions to problem (1) given by

$$\omega_{(\alpha,\theta,\beta)}(x) = \frac{N}{\alpha} f^* \left(\theta + \frac{x - x_0}{N} \alpha \right) - \frac{N}{\alpha} f^*(\theta) + \beta.$$

A remarkable feature of this class of solutions to problem 1 is provided by the following result:

THEOREM 2. *Let f be an extended valued, convex, lower semicontinuous function with superlinear growth. Then*

(a) *when f^* is differentiable at θ, as α tends to 0, the function $\omega_{(\alpha,\theta,\beta)}$ converges to the affine map $\langle \nabla f^*(\theta), x - x_0 \rangle + \beta$, a solution to problem (4).*

(b) *in general, as α tends to 0^+, the function $\omega_{(\alpha,\theta,\beta)}$ converges to $h^+_{\theta,x_0,\beta}$, the solution to problem (4), presented in Theorem 1 of [1]; as α tends to 0^-, $\omega_{(\alpha,\theta,\beta)}$ converges to $h^-_{\theta,x_0,\beta}$.*

Proof.

(a) From the assumption of differentiability we have that

$$\lim_{\alpha \to 0} \frac{N}{\alpha} \left[f^* \left(\theta + \frac{x - x_0}{N} \alpha \right) - f^*(\theta) \right]$$
$$= \langle \nabla f^*(\theta), x - x_0 \rangle .$$

(b) By assumption, $\partial f^*(\theta)$ is nonempty, so by Theorem 23.4 of [3], we have

$$\lim_{\alpha \to 0^+} \frac{N}{\alpha} f^* \left(\theta + \frac{x - x_0}{N} \alpha \right) - \frac{N}{\alpha} f^*(\theta)) = \sup_{k \in \partial f^*(\theta)} \langle k, x - x_0 \rangle .$$

By Theorem 1 of [1], the map $\sup_{k\in\partial f^*(\theta)}\langle k, x-x_0\rangle$ is a solution to (4) and the claim follows. The proof is analogous for the second claim, taking into account that the function $\omega_{(\alpha,\theta,\beta)}$ becomes concave in this case. □

Example. Consider the problem

$$\text{minimize}\int_\Omega [G(u'(x))+\alpha u(x)]\,dx,$$

where $\alpha>0$ and G is

$$G(\xi)=\begin{cases}\sqrt{2}\|\xi\| & \text{if } \|\xi\|\le\sqrt{2},\\ 1+\frac{1}{2}\|\xi\|^2 & \text{if } \|\xi\|\ge\sqrt{2}.\end{cases} \tag{5}$$

We have

$$G^*(p)=\begin{cases}0 & \text{if } \|p\|\le\sqrt{2},\\ \frac{1}{2}\|p\|^2-1 & \text{if } \|p\|\ge\sqrt{2}.\end{cases}$$

In particular, for $\|\theta\|=\sqrt{2}$, so that $G^*(\theta)=0$, and $x_0=0$, we obtain

$$\omega_{(\alpha,\theta,0)}=\begin{cases}0 & \text{if } \|\theta+\frac{\alpha}{N}x\|\le\sqrt{2},\\ \frac{N}{\alpha}\left(\frac{1}{2}\|\theta+\frac{\alpha}{N}x\|^2-1\right) & \text{if } \|\theta+\frac{\alpha}{N}x\|\ge\sqrt{2}.\end{cases}$$

As $\alpha\downarrow 0$, the set of points $\{x:\|\theta+\frac{\alpha}{N}x\|\le\sqrt{2}\}$ grows to the half space $\{x:\langle x,\theta\rangle\le 2\}$, and $\omega_{(\alpha,\theta,0)}$ converges pointwise to

$$\begin{cases}0 & \text{if } \langle x,\theta\rangle\le 2,\\ \langle\theta,x\rangle & \text{if } \langle x,\theta\rangle\ge 2.\end{cases}$$

This last function is

$$\sup_{k\in\partial G^*(\theta)}\{\langle k,x\rangle\}=\left(I_{\partial G^*(\theta)}\right)^*(x),$$

a solution to (4).

REFERENCES

[1] A. Cellina, *Comparison Results and Estimates on the Gradient without Strict Convexity*, SIAM J. Control Optim., to appear.

[2] I. Ekeland and R. Temam, *Analyse convexe et problémes variationnels*, Gauthier-Villars, Paris-Brussels-Montreal, 1974.

[3] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1972.

SIAM J. OPTIM.
Vol. 18, No. 3, pp. 717–732

METRIC REGULARITY IN CONVEX SEMI-INFINITE OPTIMIZATION UNDER CANONICAL PERTURBATIONS*

M. J. CÁNOVAS†, D. KLATTE‡, M. A. LÓPEZ§, AND J. PARRA†

Abstract. This paper is concerned with the Lipschitzian behavior of the optimal set of convex semi-infinite optimization problems under continuous perturbations of the right-hand side of the constraints and linear perturbations of the objective function. In this framework we provide a sufficient condition for the metric regularity of the inverse of the optimal set mapping. This condition consists of the Slater constraint qualification, together with a certain additional requirement in the Karush–Kuhn–Tucker conditions. For linear problems this sufficient condition turns out to be also necessary for the metric regularity, and it is equivalent to some well-known stability concepts.

Key words. metric regularity, optimal set, Lipschitz properties, semi-infinite programming, convex programming

AMS subject classifications. 90C34, 49J53, 90C25, 90C31, 90C05

DOI. 10.1137/060658345

1. Introduction.

We consider the canonically perturbed convex semi-infinite programming problem, in $\mathbb{R}^n$,

$$P(c,b): \quad \begin{array}{l} \text{Inf } f(x) + c'x \\ \text{s.t. } g_t(x) \le b_t, \ \ t \in T, \end{array} \tag{1}$$

where $x \in \mathbb{R}^n$ is the vector of decision variables, regarded as a column-vector, $c \in \mathbb{R}^n$, c' denotes the transpose of c, the index set T is a compact metric space, $f : \mathbb{R}^n \to \mathbb{R}$ and $g_t : \mathbb{R}^n \to \mathbb{R}$, $t \in T$, are given convex functions in such a way that $(t,x) \mapsto g_t(x)$ is continuous on $T \times \mathbb{R}^n$, and $b \in C(T,\mathbb{R})$, i.e., $T \ni t \mapsto b_t \in \mathbb{R}$ is continuous on T.

In this setting, the pair $(c,b) \in \mathbb{R}^n \times C(T,\mathbb{R})$ is regarded as the parameter to be perturbed. We denote by $\sigma(b)$ the constraint system associated with $P(c,b)$, i.e.,

$$\sigma(b) := \{ g_t(x) \le b_t, \ \ t \in T \}.$$

The parameter space $\mathbb{R}^n \times C(T,\mathbb{R})$ is endowed with the norm

$$\|(c,b)\| := \max \{ \|c\|, \|b\|_\infty \}, \tag{2}$$

where $\mathbb{R}^n$ is equipped with any given norm $\|\cdot\|$ and $\|b\|_\infty := \max_{t\in T} |b_t|$. The corresponding dual norm in $\mathbb{R}^n$ is given by $\|u\|_* := \max \{u'x \mid \|x\| \le 1\}$.

*Received by the editors April 27, 2006; accepted for publication (in revised form) October 27, 2006; published electronically October 4, 2007. This work was partially supported by grants MTM2005-08572-C03 (01-02) from MEC (Spain) and FEDER (E.U.), and ACOMP06/117-203 from Generalitat Valenciana (Spain).

http://www.siam.org/journals/siopt/18-3/65834.html

†Operations Research Center, Miguel Hernández University of Elche, 03202 Elche (Alicante), Spain (canovas@umh.es, parra@umh.es).

‡Institut für Operations Research, Universität Zürich, Moussonstrasse 15, CH-8044 Zürich, Switzerland (klatte@ior.unizh.ch).

§Department of Statistics and Operations Research, University of Alicante, 03071 Alicante, Spain (marco.antonio@ua.es).

Associated with the parametric family of problems $P(c,b)$, we consider the set-valued mappings $\mathcal{G}:\mathbb{R}^n \rightrightarrows C(T,\mathbb{R})$ and $\mathcal{G}^*:\mathbb{R}^n \rightrightarrows \mathbb{R}^n \times C(T,\mathbb{R})$ given by

$$\mathcal{G}(x) := \{b \in C(T,\mathbb{R}) \mid g_t(x) \leq b_t \text{ for all } t \in T\},$$

$$\mathcal{G}^*(x) := \left\{(c,b) \in \mathbb{R}^n \times \mathcal{G}(x) \mid x \in \arg\min\left\{f(y) + c'y \mid y \in \mathcal{G}^{-1}(b)\right\}\right\}.$$

The corresponding inverse mappings will be denoted by $\mathcal{F}$ and $\mathcal{F}^*$, respectively. Observe that $\mathcal{F}(b)$ and $\mathcal{F}^*(c,b)$ are, respectively, the feasible set and the optimal set (set of optimal solutions) of $P(c,b)$, i.e.,

$$\mathcal{F}(b) := \{x \in \mathbb{R}^n \mid g_t(x) \leq b_t \text{ for all } t \in T\},$$

$$\mathcal{F}^*(c,b) := \arg\min\{f(x) + c'x \mid x \in \mathcal{F}(b)\}.$$

Finally, by Π_c and Π_s we denote the sets of parameters corresponding to consistent or solvable problems, respectively; i.e.,

$$\Pi_c := \{(c,b) \in \mathbb{R}^n \times C(T,\mathbb{R}) \mid \mathcal{F}(b) \neq \emptyset\}$$

and

$$\Pi_s := \{(c,b) \in \mathbb{R}^n \times C(T,\mathbb{R}) \mid \mathcal{F}^*(c,b) \neq \emptyset\}.$$

According to Corollary 8.3.3 and Theorem 8.7 in [24], if $\sigma(b)$ and $\sigma(b^1)$ are both consistent, $\mathcal{F}(b)$ and $\mathcal{F}(b^1)$ have the same recession cone.

This paper is concerned with the metric regularity of $\mathcal{G}^*$ at a given $\overline{x}$ for $(\overline{c},\overline{b}) \in \mathcal{G}^*(\overline{x})$, that is, with the existence of neighborhoods U of $\overline{x}$ and V of $(\overline{c},\overline{b})$ and a constant $\kappa \geq 0$ such that

$$d(x,\mathcal{F}^*(c,b)) \leq \kappa d((c,b),\mathcal{G}^*(x)) \text{ for all } x \in U \text{ and all } (c,b) \in V, \tag{3}$$

where, as usual, $d(x,\emptyset) = +\infty$. In section 3 we provide a sufficient condition, (10), for this property. Essentially, it is a Karush–Kuhn–Tucker (KKT) type condition with some additional requirements.

In the particular case of linear problems of the form

$$\begin{aligned} P(c,b): \quad & \operatorname{Inf} c'x \\ & \text{s.t. } a_t'x \geq b_t, \ \ t \in T, \end{aligned} \tag{4}$$

where $a \in C(T,\mathbb{R}^n)$ is a given function, this algebraic condition is given by (9), and it turns out to be equivalent to a condition introduced by Nürnberger [22, Condition (2) in Thm. 1.4], in relation to the stability of the strong uniqueness of minimizers (see also [11] and [13], dealing with linear optimization problems without continuity assumptions). Moreover in the linear setting, the referred condition is not only sufficient but also necessary for the metric regularity of $\mathcal{G}^*$ at $\overline{x}$ for $(\overline{c},\overline{b})$.

The metric regularity is a basic quantitative property of mappings in variational analysis which is widely used in both theoretical and computational studies. In order to illustrate how this concept works in our context, let $\overline{x}$ be an optimal solution of $P(\overline{c},\overline{b})$ and let (c_a,b_a) and x_a be close enough approximations to $(\overline{c},\overline{b})$ and $\overline{x}$, respectively. Then problem $P(c_a,b_a)$ has an optimal solution whose distance to x_a is bounded by κ times $d((c_a,b_a),\mathcal{G}^*(x_a))$. The latter distance is usually easy to compute

or estimate, while finding an exact solution of $P(c_a, b_a)$ might be considerably difficult. For instance, a possible choice of parameters which make x_a optimal is $c = \bar{c}$ and b such that x_a is feasible for $\sigma(b)$ and some suitably chosen constraints are active at x_a (according to the KKT condition). See section 3 for details. The metric regularity of a set-valued mapping turns out to be equivalent to the pseudo-Lipschitz property, also called the Aubin property, of the inverse mapping (see, for instance, [19], [25] and the references therein). Specifically, the Aubin property in our context reads as follows: There exist neighborhoods U of $\overline{x}$ and V of $\left(\bar{c}, \bar{b}\right)$ and a constant $\kappa \geq 0$ such that

$$d\left(x^2, \mathcal{F}^*\left(c^1, b^1\right)\right) \leq \kappa d\left(\left(c^1, b^1\right), \left(c^2, b^2\right)\right) \tag{5}$$

for all $\left(c^1, b^1\right), \left(c^2, b^2\right) \in V$ and all $x^2 \in U \cap \mathcal{F}^*\left(c^2, b^2\right)$. Other Lipschitz/regularity properties also can be traced back to [19], [25].

In our context of problems (1), the metric regularity of $\mathcal{G}^*$ (i.e., the pseudo-Lipschitz property of $\mathcal{F}^*$) at a point of its graph is equivalent to the *strong Lipschitz stability* of $\mathcal{F}^*$ (see Lemma 5), which reads as follows: There exist open neighborhoods U of $\overline{x}$ and V of $\left(\bar{c}, \bar{b}\right)$ and a constant $\kappa \geq 0$ such that $\mathcal{F}^*(c, b) \cap U$ is a singleton, $\{x(c, b)\}$, for all $(c, b) \in V$ and

$$\left\|x\left(c^1, b^1\right) - x\left(c^2, b^2\right)\right\| \leq \kappa \left\|\left(c^1, b^1\right) - \left(c^2, b^2\right)\right\| \text{ for all } \left(c^1, b^1\right), \left(c^2, b^2\right) \in V.$$

Note that because of the convexity of $\mathcal{F}^*(c, b)$, we already have $\mathcal{F}^*(c, b) = \{x(c, b)\}$ for all $(c, b) \in V$. In other words, the *strong Lipschitz stability* of $\mathcal{F}^*$ at $\left(\left(\bar{c}, \bar{b}\right), \overline{x}\right)$ is equivalent to the local single-valuedness and Lipschitz continuity of $\mathcal{F}^*$ near $\left(\left(\bar{c}, \bar{b}\right), \overline{x}\right)$ [17], [19], [26]. The fact that the pseudo-Lipschitz property of the global optimal solution set mapping $\mathcal{S}$ of a parametric optimization problem implies strong Lipschitz stability of $\mathcal{S}$ holds for a rather general class of optimization problems (see again Lemma 5). In the particular case of linear problems, we can add as a third equivalent property the *local single-valuedness and continuity* of $\mathcal{F}^*$ (a Kojima-type stability condition under specific perturbations [21], [26]).

Section 5.3 in [20] clarifies the relationship between the strong Lipschitz stability and the strong Kojima stability. Specifically, as a straightforward consequence of Corollary 5.5 there, one obtains the equivalence between these two properties when applied to finite linear optimization problems. In this way, Theorem 16 below, confined to the linear case, extends the fulfillment of these equivalences to the case of infinitely many constraints.

That paper [20] was concerned with the strong Lipschitz stability of the stationary solution map (in the KKT sense) in our context of problems (1), with T finite, where the functions included in the model are assumed to belong to the class $C^{1,1}$, and under the general assumption of the Mangasarian–Fromowitz constraint qualification (MFCQ). The more general case in which the functions f and g also depend on a parameter $\tau \in \mathcal{T} \subset \mathbb{R}^r$ is dealt with in [19, sect. 8]. Note that if the constraint functions g_t of the convex semi-infinite problem (1) are differentiable, then the (extended) MFCQ is nothing else but the Slater CQ (i.e., the existence of a strict solution of the associated constraint system). The fulfillment of both the Slater condition and the boundedness (and nonemptiness) of the set of optimal solutions yields high stability for optimization problems in different frameworks (see, for instance, [18, Thm. 1] and [5, Thm. 4.2] in relation to the Lipschitz continuity of the optimal value).

There are different contributions to the stability theory for the feasible and the optimal set mappings in linear semi-infinite optimization. The article [10] analyzed

the (Berge) lower semicontinuity of the feasible set mapping $\mathcal{F}$ in the more general context in which there is no continuity assumption and the parameters are $(a,b) \in (\mathbb{R}^n \times \mathbb{R})^T$, the latter being endowed with an appropriate extended distance. On the other hand, the lower and upper semicontinuity of $\mathcal{F}^*$ in the general context of parameters $(c,(a,b)) \in \mathbb{R}^n \times (\mathbb{R}^n \times \mathbb{R})^T$ were analyzed in [5] in the linear case, and in [8] in the convex case. More details about stability of linear semi-infinite problems and their constraint systems in this general context (no continuity assumption) are gathered in [9, Chapters 6 and 10]. The *continuous case*, in which T is a compact Hausdorff space, the functions a and b are continuous on T, and all the parameters may be (continuously) perturbed, was analyzed, e.g., in [3] and [7]. Note also that classical parametric optimization (see, e.g., [1], [2], [16]) applies to this and more general settings by writing the constraints as one aggregated inequality, like $\max_{t\in T}(g_t(x) - b_t) \leq 0$ in the case of (1). In the current context of continuous perturbations of only the right-hand side of the system, the metric regularity of the mapping $\mathcal{G}$, in the linear case, was approached in [4].

Next, we summarize the structure of the paper. Section 2 gathers some preliminaries about convex analysis and multifunctions. Moreover we include here some results about the stability of $\mathcal{F}$ and its relation with continuity properties of $\mathcal{F}^*$. Specifically, Lemma 3 shows the equivalence among some relevant stability criteria concerning the feasible set. Proposition 4 provides a sufficient condition for the lower semicontinuity of $\mathcal{F}^*$, which constitutes a key step in the analysis of the metric regularity of $\mathcal{G}^*$. In section 3 we introduce, after some motivation, condition (10). Some consequences of this condition are gathered in Proposition 9. Theorem 10 shows that condition (10) is sufficient for the metric regularity of $\mathcal{G}^*$ in the convex case. Section 4 deals with the linear case. Theorem 16 establishes the equivalence between the specification of (10) for the linear case and several well-known stability concepts concerning the optimal set, including the metric regularity of $\mathcal{G}^*$. Finally, section 5 shows at a glance the main results of the paper.

2. Preliminaries and first results. In this section we provide further notation and some preliminary results. Given $X \subset \mathbb{R}^k$, $k \in \mathbb{N}$, we denote by $conv\,(X)$ and $cone\,(X)$ the convex hull and the conical convex hull of X, respectively. We assume that $cone\,(X)$ always contains the zero vector of $\mathbb{R}^k$, 0_k. We shall also assume $conv\,(\emptyset) = \emptyset$ and $cone\,(\emptyset) := \{0_k\}$. If X is a closed convex set, $O^+(X)$ represents the recession cone of X.

If X is a subset of any topological space, $\mathrm{int}\,(X)$ and $\mathrm{cl}(X)$ will represent the interior and the closure of X, respectively. A typical element of $cone\,(\{x_i, i \in I\})$, where I is any index set, is represented as $\sum_{i\in I} \lambda_i x_i$, where $\lambda = (\lambda_i)_{i\in I}$ belongs to the cone $\mathbb{R}_+^{(I)}$ of all functions from I to $\mathbb{R}_+ := [0, +\infty[$ with finite support, i.e., taking positive values at only finitely many points of I. Generically, sequences will be indexed by $r \in \mathbb{N}$, and $\lim_r$ should be interpreted as $\lim_{r\to\infty}$.

Let $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper closed convex function. By $\partial h\,(x)$ we denote the *subdifferential* of h at x, and by $h0^+$ the *recession function* of h, i.e., the sublinear function whose epigraph is the recession cone of the epigraph of h.

Observe that our problem $P(c,b)$ is equivalent to the unconstrained problem

$$\operatorname*{Inf}_{x\in\mathbb{R}^n} \left\{h(x) := f(x) + c'x + \delta_{\mathcal{F}(b)}(x)\right\}, \tag{6}$$

where $\delta_{\mathcal{F}(b)}$ is the indicator function of $\mathcal{F}(b)$ (i.e., $\delta_{\mathcal{F}(b)}(x) = 0$ if $x \in \mathcal{F}(b)$, and $\delta_{\mathcal{F}(b)}(x) = +\infty$ if $x \notin \mathcal{F}(b)$). We shall use the recession function of h, which, thanks

to [24, Thm. 9.3], turns out to be

$$h0^+(y) = f0^+(y) + c'y + \delta_{\mathcal{F}(b)}0^+(y)$$
$$= f0^+(y) + c'y + \delta_{O^+(\mathcal{F}(b))}(y).$$

Associated with problem (1), for each $x \in \mathcal{F}(b)$ we consider

$$T_b(x) = \{t \in T \mid g_t(x) = b_t\} \quad \text{and} \quad A_b(x) = cone\left(\bigcup_{t\in T_b(x)} (-\partial g_t(x))\right).$$

Recall that $A_b(x) = \{0_n\}$ if $T_b(x) = \emptyset$. For our model (1), $\sigma(b)$ satisfies the *Slater condition* if $T_b(x^0)$ is empty for some $x^0 \in \mathcal{F}(b)$, in which case x^0 is referred to as a Slater point of $\sigma(b)$ (see [9, sect. 7.5]). Note that the continuity of $t \mapsto g_t\left(x^0\right)$ together with the compactness of T entails that x^0 is a Slater point of $\sigma(b)$ if and only if there exists some slack $\rho > 0$ such that $g_t\left(x^0\right) \leq b_t - \rho$ for all $t \in T$.

LEMMA 1. *Let $(c, b) \in \mathbb{R}^n \times C(T, \mathbb{R})$ and $x \in \mathbb{R}^n$. One has the following for the parametric problem* (1):

(i) *If $\sigma(b)$ satisfies the Slater condition, then $A_b(x)$ is closed* [9, Thm. 7.9].

(ii) *KKT conditions* (see [9, (7.9) and Thm. 7.8]): *If $x \in \mathcal{F}(b)$ and $(c + \partial f(x)) \cap A_b(x) \neq \emptyset$, then $x \in \mathcal{F}^*(c, b)$. The converse holds when $\sigma(b)$ satisfies the Slater condition.*

Next we recall some well-known continuity concepts for set-valued mappings. If $\mathcal{Y}$ and $\mathcal{Z}$ are two metric spaces and $\mathcal{H} : \mathcal{Y} \rightrightarrows \mathcal{Z}$ is a set-valued mapping, $\mathcal{H}$ is said to be *lower semicontinuous* (lsc, in brief), in the classical sense of Berge, at $y \in \mathcal{Y}$ if, for each open set $W \subset \mathcal{Z}$ such that $W \cap \mathcal{H}(y) \neq \emptyset$, there exists an open set $U \subset \mathcal{Y}$, containing y, such that $W \cap \mathcal{H}(y^1) \neq \emptyset$ for each $y^1 \in U$. The mapping $\mathcal{H}$ is *upper semicontinuous* (usc, for short), in the sense of Berge, at $y \in \mathcal{Y}$ if, for each open set $W \subset \mathcal{Z}$ such that $\mathcal{H}(y) \subset W$, there exists an open neighborhood of y in $\mathcal{Y}$, U, such that $\mathcal{H}(y^1) \subset W$ for every $y^1 \in U$. We say that $\mathcal{H}$ is *closed* at $y \in \mathcal{Y}$ if for all sequences $\{y^r\} \subset \mathcal{Y}$ and $\{z^r\} \subset \mathcal{Z}$ satisfying $\lim_{r\to\infty} y^r = y$, $\lim_{r\to\infty} z^r = z$, and $z^r \in \mathcal{H}(y^r)$, one has $z \in \mathcal{H}(y)$. Obviously, $\mathcal{H}$ is closed on $\mathcal{Y}$ (at every point $y \in \mathcal{Y}$) if the graph of $\mathcal{H}$, $gph(\mathcal{H}) := \{(y, z) \in \mathcal{Y} \times \mathcal{Z} : z \in \mathcal{H}(y)\}$, is closed (in the product topology). In what follows, $rge(\mathcal{H})$ will represent the image set of $\mathcal{H}$.

The following property of our optimal set mapping $\mathcal{F}^*$ is a straightforward consequence of [1, Thm. 4.3.3] and will be used later on.

LEMMA 2. *Let $(\bar{c}, \bar{b}) \in \mathbb{R}^n \times C(T, \mathbb{R})$. Assume that $\mathcal{F}$ is lsc at $\bar{b}$ and $\mathcal{F}^*(\bar{c}, \bar{b})$ is nonempty and bounded. Then $\mathcal{F}^*$ is usc at $(\bar{c}, \bar{b})$.*

Note that our mapping $\mathcal{F}$ is closed on $C(T, \mathbb{R})$ due to the continuity of each g_t. The lower semicontinuity of $\mathcal{F}$ turns out to be equivalent to other stability properties referred above (see [12] for a discussion about conditions (i)–(iii) in the following lemma).

LEMMA 3. (See [4, Thm. 2.1] for the linear case with equality/inequality constraints.) *Let $\bar{b} \in rge(\mathcal{G})$. The following statements are equivalent:*

(i) *$\sigma\left(\bar{b}\right)$ satisfies the Slater condition.*

(ii) *$\mathcal{F}$ is lsc at $\bar{b}$.*

(iii) $\bar{b} \in \text{int}(rge(\mathcal{G}))$.

(iv) *$\mathcal{G}$ is metrically regular at any $x \in \mathcal{F}\left(\bar{b}\right)$ for $\bar{b}$.*

(v)

$$0_n \notin conv\left(\bigcup_{t\in T_b(x)} \partial g_t(x)\right) \text{ for all } x \in \mathcal{F}\left(\overline{b}\right) \text{ such that } T_{\overline{b}}(x) \neq \emptyset. \tag{7}$$

Proof. (i)⇒(ii). Define the function $G : \mathbb{R}^n \times C(T,\mathbb{R}) \to \mathbb{R}$ by

$$G(x,b) := \max_{t\in T} (g_t(x) - b_t).$$

Hence $\mathcal{F}(b) = \{x \mid G(x,b) \leq 0\}$. By classical parametric optimization (cf., e.g., [2], [16]), G is continuous, since $(t,x,b) \mapsto g_t(x) - b_t$ is continuous and T is nonempty and compact. Obviously, for given b, $G(\cdot,b)$ is convex. Since for $x \in \mathcal{F}\left(\overline{b}\right)$ we have $G(x,\overline{b}) = 0$ if and only if $T_{\overline{b}}(x) \neq \emptyset$, statement (i) is equivalent to the existence of $\overline{x}$ such that $G(\overline{x},\overline{b}) < 0$. Now, Theorem 12 in [16] applies.

(ii)⇒(iii). It comes straightforwardly from the definitions, taking into account that (iii) may be interpreted as $\sigma\left(b^1\right)$ being consistent ($\mathcal{F}\left(b^1\right) \neq \emptyset$) for all b^1 in some neighborhood of $\overline{b}$.

(iii)⇒(i). It follows from the following fact: For $\varepsilon > 0$ small enough, $\mathcal{F}\left(b^\varepsilon\right) \neq \emptyset$, where $b^\varepsilon \in C\left(T,\mathbb{R}\right)$ is given by $b_t^\varepsilon := \overline{b}_t - \varepsilon$, $t \in T$. In this case, any feasible point of $\sigma\left(b^\varepsilon\right)$ is a Slater point of $\sigma\left(\overline{b}\right)$ with slack ε.

(iii)⇔(iv). This equivalence is established via the Robinson–Ursescu theorem (see, for instance, [6]) for mappings between Banach spaces having a closed convex graph. We have already mentioned that $gph(\mathcal{G})$ is closed, and it is also convex, due to the convexity of each g_t.

(i)⇔(v). With G as above, let $g(x) := G(x,\overline{b})$. Thus, (i) equivalently means that $g(\overline{x}) < 0$ is satisfied for some $\overline{x}$, which holds if and only if every point $x \in \mathcal{F}\left(\overline{b}\right)$ such that $T_{\overline{b}}(x) \neq \emptyset$ is not a minimum of g. By [15, Thm. VI.4.4.2], the latter is equivalent to the following fact: For every point $x \in \mathcal{F}\left(\overline{b}\right)$ such that $T_{\overline{b}}(x) \neq \emptyset$ we have

$$0_n \notin \partial g(x) = conv\left(\bigcup_{t\in T_b(x)} \partial g_t(x)\right),$$

and this is precisely (v). □

The following proposition accounts for some properties of $\mathcal{F}^*$ in relation to $\mathcal{F}$ (see also Lemma 2).

PROPOSITION 4. (i) *If* $(\overline{c},\overline{b}) \in \mathrm{int}(\Pi_s)$*, then* $\mathcal{F}^*(\overline{c},\overline{b})$ *is a nonempty bounded set.*

(ii) *Assume that* $(\overline{c},\overline{b}) \in \mathrm{int}(\Pi_c)$ *and that* $\mathcal{F}^*(\overline{c},\overline{b})$ *is a nonempty bounded set. Then* $(\overline{c},\overline{b}) \in \mathrm{int}(\Pi_s)$ *and* $\mathcal{F}^*(c,b)$ *is also a nonempty bounded set for* (c,b) *in a certain neighborhood of* $(\overline{c},\overline{b})$.

(iii) *If* $\mathcal{F}$ *is lsc at* $\overline{b}$*, then* $\mathcal{F}^*$ *is closed at* $\left(\overline{c},\overline{b}\right)$.

(iv) *If* $\mathcal{F}$ *is lsc at* $\overline{b}$ *and* $\mathcal{F}^*\left(\overline{c},\overline{b}\right)$ *is a singleton, then* $\mathcal{F}^*$ *is lsc at* $\left(\overline{c},\overline{b}\right)$.

Proof. (i) Let $(\overline{c},\overline{b}) \in \mathrm{int}(\Pi_s)$, and assume that $\mathcal{F}^*(\overline{c},\overline{b})$ is unbounded. Take $u \in O^+(\mathcal{F}^*(\overline{c},\overline{b}))$, $u'u = 1$, and consider the sequence in Π_c, $\left(\overline{c} - \frac{1}{r}u, \overline{b}\right)$, $r = 1,2,\ldots$, which obviously converges to $(\overline{c},\overline{b})$. Now, for $\lambda \geq 0$ and $\overline{x} \in \mathcal{F}^*(\overline{c},\overline{b}) \subset \mathcal{F}(\overline{b})$, and representing by v the optimal value of $P(\overline{c},\overline{b})$, we have

$$f(\overline{x} + \lambda u) + \left(\overline{c} - \frac{1}{r}u\right)'(\overline{x} + \lambda u) = v - \frac{1}{r}u'\overline{x} - \frac{\lambda}{r}.$$

By letting $\lambda \to +\infty$, it follows that the objective function of $P\left(\overline{c} - \frac{1}{r}u, \overline{b}\right)$ is unbounded from below, and this contradicts the assumption $(\overline{c}, \overline{b}) \in \operatorname{int}(\Pi_s)$.

(ii) Since $\mathcal{F}^*(\overline{c}, \overline{b})$ is nonempty and bounded, [15, Prop. IV.3.2.5] yields $\overline{h}0^+(y) > 0$ for all $y \neq 0_n$, where $\overline{h}$ is the function introduced in (6), associated to the nominal parameter $(\overline{c}, \overline{b})$. Since $\overline{h}0^+$ is lsc

$$\varepsilon := \min\{\overline{h}0^+(y) \mid \|y\|_* = 1\} > 0.$$

Consider any parameter (c, b) such that $\|c - \overline{c}\| < \varepsilon$ and that is close enough to $(\overline{c}, \overline{b})$ to be sure that $(c, b) \in \Pi_c$. If h is the associated function (see (6)) and $\|y\|_* = 1$, we can write

$$\begin{aligned} h0^+(y) &= f0^+(y) + c'y + \delta_{O^+(\mathcal{F}(b))}(y) \\ &= f0^+(y) + \overline{c}'y + \delta_{O^+(\mathcal{F}(\overline{b}))}(y) + (c - \overline{c})'y \\ &= \overline{h}0^+(y) + (c - \overline{c})'y \qquad (8) \\ &\geq \overline{h}0^+(y) - \|c - \overline{c}\| \\ &> \overline{h}0^+(y) - \varepsilon \geq 0. \end{aligned}$$

Since (8) entails $h0^+(y) > 0$ for all $y \neq 0_n$, [15, Prop. IV.3.2.5] implies that $\mathcal{F}^*(c, b)$ is a nonempty bounded set.

(iii) Since $\mathcal{F}$ is closed at $\overline{b}$, this is a classical result; see [16, Thm. 8].

(iv) Since $\mathcal{F}^*\left(\overline{c}, \overline{b}\right)$ is a singleton, it holds by definition that $\mathcal{F}^*$ is lsc at $\left(\overline{c}, \overline{b}\right)$ if $\mathcal{F}^*$ is both usc at $\left(\overline{c}, \overline{b}\right)$ and nonempty-valued near $\left(\overline{c}, \overline{b}\right)$. The first property follows from Lemma 2, the second one from Corollary 9.1 in [16]. ∎

Problem (1) fits into the more general class of parametric problems given by

$$\begin{aligned} \mathcal{P}(c, b): \quad & \operatorname{Inf} f(x) + c'x \\ & \text{s.t. } x \in \mathcal{M}(b), \end{aligned}$$

where f is any real-valued function defined on $\mathbb{R}^n$, $\mathcal{M}$ is any multifunction which maps a metric space $\mathcal{Y}$ to $\mathbb{R}^n$, and $(c, b) \in \mathbb{R}^n \times \mathcal{Y}$ varies in some neighborhood of $(\overline{c}, \overline{b}) \in \mathbb{R}^n \times \mathcal{Y}$. If we define

$$\mathcal{F}^*(c, b) := \arg\min\{f(x) + c'x \mid x \in \mathcal{M}(b)\},$$

we obtain the following result without any assumption about continuity.

LEMMA 5 (Corollary 4.7 in [19]). *Let $((\overline{c}, \overline{b}), \overline{x}) \in gph(\mathcal{F}^*)$. Then $\mathcal{F}^*$ is pseudo-Lipschitz at $((\overline{c}, \overline{b}), \overline{x})$ if and only if $\mathcal{F}^*$ is strongly Lipschitz stable at this point.*

Proof. To show the nontrivial direction, let $\mathcal{F}^*$ be pseudo-Lipschitz at $((\overline{c}, \overline{b}), \overline{x})$. Hence, by Corollary 4.7 in [19], $\mathcal{F}^*(c, b)$ is a singleton for (c, b) near $(\overline{c}, \overline{b})$. This implies strong Lipschitz stability at (and hence, by definition of that stability, *near*) $((\overline{c}, \overline{b}), \overline{x})$. ∎

3. A sufficient condition for the metric regularity of $\mathcal{G}^*$. This section provides a KKT-type condition which is sufficient for the metric regularity $\mathcal{G}^*$ at $\overline{x}$ for $\left(\overline{c}, \overline{b}\right) \in \mathcal{G}^*(\overline{x})$ in the context of convex problems (1). The relationship between this condition and the strong uniqueness of optimal solutions is explored, too. The specification of this KKT-type property for linear problems (4) turns out to be also

necessary for the metric regularity. The next example partially motivates this algebraic condition in the linear case.

Example 6. Consider the problem, in $\mathbb{R}^2$ (with the Euclidean norm),

$$P\left(\bar{c},\bar{b}\right) := \operatorname{Inf}\left\{x_1 \mid x_1 - x_2 \geq 0,\, x_1 + x_2 \geq 0,\, x_1 \geq 0\right\}.$$

Here $\bar{c} = (1,0)'$ and $\bar{b} = 0_3$.

One has $\mathcal{F}^*\left(\bar{c},\bar{b}\right) = \{0_2\}$. If we consider the perturbed problem $P\left(c^r,b^r\right)$, with $b^r := (0,0,1/r)'$ and $c^r = \left(1,-1/r^2\right)$, we have $\mathcal{F}^*\left(c^r,b^r\right) = \left\{\left(\frac{1}{r},\frac{1}{r}\right)\right\}$. So, by taking $x^r = \left(\frac{1}{r},0\right)'$, we obtain

$$d\left(x^r,\mathcal{F}^*\left(c^r,b^r\right)\right) = \frac{1}{r} \quad \text{and} \quad d\left(\left(c^r,b^r\right),\mathcal{G}^*\left(x^r\right)\right) \leq d\left(\left(c^r,b^r\right),\left(\bar{c},b^r\right)\right) = \frac{1}{r^2}.$$

Hence, $d\left(x^r,\mathcal{F}^*\left(c^r,b^r\right)\right) \geq r d\left(\left(c^r,b^r\right),\mathcal{G}^*\left(x^r\right)\right)$, $r = 1,2,\ldots$. Therefore, $\mathcal{G}^*$ is not metrically regular at 0_2 for $\left(\bar{c},\bar{b}\right)$.

The key fact in this example is that $\bar{c}$ belongs to the convex cone generated by *one* vector, associated with the active constraints in $\overline{x}$, in the *two*-dimensional Euclidean space. The following property, referred to as a given $\left(\overline{x},\left(\bar{c},\bar{b}\right)\right) \in gph(\mathcal{G}^*)$ in the linear case (4), avoids the previous situation (here $|D|$ denotes the cardinality of D):

$$\begin{array}{l} \sigma\left(\bar{b}\right) \text{ satisfies the Slater condition and there is no } D \subset T_{\bar{b}}\left(\overline{x}\right) \\ \text{with } |D| < n \text{ such that } \bar{c} \in cone\left(\{a_t,\ t \in D\}\right). \end{array} \tag{9}$$

The following natural extension of (9) for the convex problem (1) will play a crucial role in this section; in fact, it constitutes the announced sufficient condition for the metric regularity of $\mathcal{G}^*$ at $\left(\overline{x},\left(\bar{c},\bar{b}\right)\right)$:

$$\begin{array}{l} \sigma\left(\bar{b}\right) \text{ satisfies the Slater condition and there is no } D \subset T_{\bar{b}}\left(\overline{x}\right) \\ \text{with } |D| < n \text{ such that } \left(\bar{c} + \partial f(\overline{x})\right) \cap cone\left(\displaystyle\bigcup_{t\in D}\left(-\partial g_t\left(\overline{x}\right)\right)\right) \neq \emptyset. \end{array} \tag{10}$$

Remark 7. Observe that condition (9) does not imply the linear independence of $\{a_t,\ t \in T_{\bar{b}}(\overline{x})\}$. Consider the example resulting from replacing the third constraint in Example 6 with any of the other two (which would appear twice in the system).

Remark 8. In the case $n = 1$, condition (10) reads as follows: $\sigma\left(\bar{b}\right)$ satisfies the Slater condition and $0 \notin \bar{c} + \partial f(\overline{x})$ (which entails $T_{\bar{b}}\left(\overline{x}\right) \neq \emptyset$).

PROPOSITION 9. *Assume that* $\left(\overline{x},\left(\bar{c},\bar{b}\right)\right) \in gph(\mathcal{G}^*)$ *verifies* (10). *Then the following conditions hold:*

(i) *There exists a neighborhood* W *of* $\left(\overline{x},\left(\bar{c},\bar{b}\right)\right)$ *such that* (10) *is satisfied when* $\left(\overline{x},\left(\bar{c},\bar{b}\right)\right)$ *is replaced by any* $(x,(c,b)) \in W \cap gph(\mathcal{G}^*)$.

(ii) *There exist* $u \in \partial f(\overline{x})$ *as well as some* $u_{t_i} \in -\partial g_{t_i}\left(\overline{x}\right)$, $t_i \in T_{\bar{b}}\left(\overline{x}\right)$, *and some* $\lambda_i > 0$ *for* $i \in \{1,\ldots,n\}$, *such that* $\{u_{t_1},\ldots,u_{t_n}\}$ *is a basis of* $\mathbb{R}^n$ *and*

$$u + \bar{c} = \sum_{i=1}^{n} \lambda_i u_{t_i}.$$

(iii) $\mathcal{F}^*\left(\bar{c},\bar{b}\right) = \{\overline{x}\}$.

(iv) $\mathcal{F}^*$ *is lsc at* $\left(\bar{c},\bar{b}\right)$.

As a consequence of the previous statements, one has the following condition:

(v) *There exists a neighborhood V of $\left(\bar{c},\bar{b}\right)$ such that $\mathcal{F}^*$ is single-valued and continuous on V.*

Proof. (i) From the equivalence (i)⇔(iii) in Lemma 3, it is clear that $\sigma(b)$ fulfills the Slater condition for b close enough to $\bar{b}$. Now, reasoning by contradiction, assume that there exists $\{(x^r,(c^r,b^r))\}\subset gph(\mathcal{G}^*)$ converging to $\left(\overline{x},\left(\bar{c},\bar{b}\right)\right)$ as well as some subgradients $u^r\in\partial f(x^r)$, $u^r_{t^r_i}\in-\partial g_{t^r_i}(x^r)$, $t^r_i\in T_{b^r}(x^r)$, $\lambda^r_i\geq 0$, $i=1,\ldots,n-1$, $r=1,2,\ldots$, such that we can write

$$u^r+c^r=\sum_{i=1}^{n-1}\lambda^r_i u^r_{t^r_i}. \tag{11}$$

In this expression we have made use of the convexity of the involved subdifferential sets.

For each $i\in\{1,\ldots,n-1\}$ the sequence $\{t^r_i\}$ has a subsequence (still denoted by $\{t^r_i\}$, for simplicity) converging to certain $\bar{t}_i\in T_{\bar{b}}(\overline{x})$, since T is compact and $g_{\bar{t}_i}(\overline{x})-\bar{b}_{\bar{t}_i}=\lim_r\left(g_{t^r_i}(x^r)-b^r_{t^r_i}\right)=0$. Let us see that the sequence $\{\gamma_r\}_{r\in\mathbb{N}}$ given by $\gamma_r:=\sum_{i=1}^{n-1}\lambda^r_i$, $r=1,2,\ldots$, must be bounded. Otherwise, we may assume without loss of generality (considering suitable subsequences) that $\lim_{r\to\infty}\gamma_r=+\infty$ and the sequence $\left\{\frac{\lambda^r_i}{\gamma_r}\right\}_{r\in\mathbb{N}}$ converges to certain $\mu_i\geq 0$ for each $i\in\{1,\ldots,n-1\}$. So, dividing by γ_r in (11) and letting $r\to+\infty$ we have (considering again appropriate subsequences of $\left\{u^r_{t^r_i}\right\}_{r\in\mathbb{N}}$ for each i)

$$\begin{aligned} &0_n=\sum_{i=1}^{n-1}\mu_i u_{\bar{t}_i},\\ &\text{with }\sum_{i=1}^{n-1}\mu_i=1\ \text{ and }\ u_{\bar{t}_i}:=\lim_r u^r_{t^r_i}\in-\partial g_{\bar{t}_i}(\overline{x}),\quad i=1,\ldots,n-1, \end{aligned} \tag{12}$$

where we have applied [24, Thm. 24.5] to sequences $\left\{g_{t^r_i}\right\}_{r\in\mathbb{N}}$, $i=1,\ldots,n-1$, and $\{x^r\}_{r\in\mathbb{N}}$ (here the continuity of $t\mapsto g_t(x)$, for all $x\in\mathbb{R}^n$, is essential to allow the use of the referred theorem). In this way we attain a contradiction with (7) in Lemma 3.

Once we have established the boundedness of $\{\gamma_r\}_{r\in\mathbb{N}}$, we may assume without loss of generality that, for each $i\in\{1,\ldots,n-1\}$, the sequence $\{\lambda^r_i\}_{r\in\mathbb{N}}$ converges to certain $\beta_i\geq 0$, $\left\{u^r_{t^r_i}\right\}_{r\in\mathbb{N}}$ converges again to certain $u_{\bar{t}_i}\in-\partial g_{\bar{t}_i}(\overline{x})$, and $\{u^r\}_{r\in\mathbb{N}}$ converges to some $u\in\partial f(\overline{x})$ (appealing again to [24, Thm. 24.5]). Thus, letting $r\to\infty$ in (11) we obtain

$$u+\bar{c}=\sum_{i=1}^{n-1}\beta_i u_{\bar{t}_i},\ \text{with }\left\{\bar{t}_1,\ldots,\bar{t}_{n-1}\right\}\subset T_{\bar{b}}(\overline{x}),$$

contradicting (10).

(ii) It follows easily from the KKT conditions (see Lemma 1), property (10), and Carathéodory's theorem.

(iii) Let $u+\bar{c}$ be represented as in (ii). If there exists $y\in\mathcal{F}^*\left(\bar{c},\bar{b}\right)\setminus\{\overline{x}\}$, then we have, by using convexity of f and taking into account

$$0\geq g_{t_i}(y)-\bar{b}_{t_i}=g_{t_i}(y)-g_{t_i}(\overline{x})\geq-u'_{t_i}(y-\overline{x})$$

as well as $\lambda_i > 0$, $i = 1, 2, \ldots, n$,

$$0 = f(y) + \bar{c}'y - f(\overline{x}) - \bar{c}'\overline{x} \geq (u + \bar{c})'(y - \overline{x})$$

$$= \sum_{i=1}^{n} \lambda_i u'_{t_i}(y - \overline{x}) \geq 0.$$

Thus, we obtain $u'_{t_i}(y - \overline{x}) = 0$ for $i = 1, \ldots, n$, contradicting the fact that $\{u_{t_1,\ldots,}u_{t_n}\}$ is a basis of $\mathbb{R}^n$.

(iv) It is a straightforward consequence of (iii) above and Proposition 4(iv) (recall also that (i)⇔(ii) in Lemma 3).

(v) Take a neighborhood $U_0 \times V_0$ of $\left(\overline{x}, \left(\bar{c}, \bar{b}\right)\right)$ contained in certain W verifying (i). Due to (iv) we may consider a neighborhood of $\left(\bar{c}, \bar{b}\right)$, say $V \subset V_0$, such that $\mathcal{F}^*(c, b) \cap U_0 \neq \emptyset$ for all $(c, b) \in V$. Now, for each $(c, b) \in V$, there exists $x \in \mathcal{F}^*(c, b) \cap U_0$, and so $(x, (c, b)) \in W \cap gph(\mathcal{G}^*)$ and (i) together with (iii) entail $\mathcal{F}^*(c, b) = \{x\}$. Finally, the continuity of the single-valued mapping $\mathcal{F}^*|_V$ comes from (i) and (iv). □

Next we present a sufficient condition for metric regularity of $\mathcal{G}^*$. By Lemma 5, the latter is equivalent to the strong Lipschitz stability of $\mathcal{F}^*$.

THEOREM 10. *For the convex semi-infinite program* (1)*, let* $\left(\overline{x}, \left(\bar{c}, \bar{b}\right)\right) \in gph(\mathcal{G}^*)$. *If condition* (10) *holds, then* $\mathcal{G}^*$ *is metrically regular at* $\overline{x}$ *for* $\left(\bar{c}, \bar{b}\right)$.

Proof. Reasoning by contradiction, assume that (10) holds, but $\mathcal{G}^*$ is not metrically regular at $\overline{x}$ for $\left(\bar{c}, \bar{b}\right)$. According to the equivalence between metric regularity of a mapping and the Aubin property of its inverse (see (5)), there must exist a sequence $\{x^r\}_{r\in\mathbb{N}} \subset \mathbb{R}^n$ converging to $\overline{x}$ and two sequences of parameters $\{(c^r, b^r)\}_{r\in\mathbb{N}}$ and $\left\{\left(\bar{c}^r, \bar{b}^r\right)\right\}_{r\in\mathbb{N}}$, both converging to $\left(\bar{c}, \bar{b}\right)$, such that, for all $r \in \mathbb{N}$, $x^r \in \mathcal{F}^*(c^r, b^r)$ and

$$d\left(x^r, \mathcal{F}^*\left(\bar{c}^r, \bar{b}^r\right)\right) > r d\left((c^r, b^r), \left(\bar{c}^r, \bar{b}^r\right)\right). \tag{13}$$

Because of condition (v) in Proposition 9 we may assume without loss of generality that, for all r, $\mathcal{F}^*\left(\bar{c}^r, \bar{b}^r\right)$ is a singleton, say $\mathcal{F}^*\left(\bar{c}^r, \bar{b}^r\right) = \{\overline{x}^r\}$. The continuity of $\mathcal{F}^*$ at $\left(\bar{c}, \bar{b}\right)$ ensures that the sequence $\{\overline{x}^r\}$ converges to $\overline{x}$ (see again Proposition 9(v)). Moreover (13) ensures, for all r, $x^r \neq \overline{x}^r$ and

$$\frac{\sup_{t\in T}\left|b_t^r - \bar{b}_t^r\right|}{\|x^r - \overline{x}^r\|} < \frac{1}{r}. \tag{14}$$

According to conditions (i) and (ii) in Proposition 9 we can write, for r large enough,

$$u^r + c^r = \sum_{i=1}^{n} \lambda_i^r u_{t_i^r}^r \quad \text{and} \quad \overline{u}^r + \bar{c}^r = \sum_{i=1}^{n} \overline{\lambda}_i^r \overline{u}_{\overline{t}_i^r}^r \tag{15}$$

for certain subgradients $u^r \in \partial f(x^r)$, $\overline{u}^r \in \partial f(\overline{x}^r)$, $u_{t_i^r}^r \in -\partial g_{t_i^r}(x^r)$, $\overline{u}_{\overline{t}_i^r}^r \in -\partial g_{\overline{t}_i^r}(\overline{x}^r)$, associated with certain indices $t_i^r \in T_{b^r}(x^r)$ and $\overline{t}_i^r \in T_{\bar{b}^r}(\overline{x}^r)$, and certain positive scalars $\lambda_i^r, \overline{\lambda}_i^r$ for $i = 1, 2, \ldots, n$. Moreover, following the same argument as in the proof of Proposition 9(i), we may assume that for each $i = 1, \ldots, n$, the sequences $\left\{\lambda_i^r\right\}_{r\in\mathbb{N}}$ and $\left\{\overline{\lambda}_i^r\right\}_{r\in\mathbb{N}}$ converge to some λ_i and $\overline{\lambda}_i$, respectively. We may also assume that, for each i, the sequences $\left\{t_i^r\right\}_{r\in\mathbb{N}}$ and $\left\{\overline{t}_i^r\right\}_{r\in\mathbb{N}}$ involved in (15) converge to t_i and $\overline{t}_i$, respectively, both belonging to $T_{\bar{b}}(\overline{x})$, and that $\{u^r\}_{r\in\mathbb{N}}$, $\{\overline{u}^r\}_{r\in\mathbb{N}}$, $\left\{u_{t_i^r}^r\right\}_{r\in\mathbb{N}}$,

and $\left\{\overline{u}^r_{t^r_i}\right\}_{r\in\mathbb{N}}$ converge to certain u, $\overline{u}\in\partial f(\overline{x})$, $u_{t_i}\in-\partial g_{t_i}(\overline{x})$, and $\overline{u}_{\bar{t}_i}\in-\partial g_{\bar{t}_i}(\overline{x})$, respectively. Thus (15) leads us to

$$u+\overline{c}=\sum_{i=1}^{n}\lambda_i u_{t_i}\quad\text{ and }\quad\overline{u}+\overline{c}=\sum_{i=1}^{n}\overline{\lambda}_i\overline{u}_{\bar{t}_i}. \tag{16}$$

Moreover, condition (10) together with Carathéodory's theorem ensures all λ_i and $\overline{\lambda}_i$ are positive and that, at the same time, $\{u_{t_1},\ldots,u_{t_n}\}$ and $\left\{\overline{u}_{\bar{t}_1},\ldots,\overline{u}_{\bar{t}_n}\right\}$ are both bases of $\mathbb{R}^n$.

On the other hand, since, for each i and each r, we have $g_{t^r_i}(x^r)=b^r_{t^r_i}$, and $g_{t^r_i}(\overline{x}^r)\le\bar{b}^r_{t^r_i}$ (recall $t^r_i\in T_{b^r}(x^r)$ and $\overline{x}^r\in\mathcal{F}(\bar{b}^r)$), we can write

$$u'_{t^r_i}\frac{x^r-\overline{x}^r}{\|x^r-\overline{x}^r\|}=-u'_{t^r_i}\frac{\overline{x}^r-x^r}{\|x^r-\overline{x}^r\|}\le\frac{g_{t^r_i}(\overline{x}^r)-g_{t^r_i}(x^r)}{\|x^r-\overline{x}^r\|}\le\frac{\bar{b}^r_{t^r_i}-b^r_{t^r_i}}{\|x^r-\overline{x}^r\|}<\frac{1}{r}, \tag{17}$$

where the last inequality comes from (14). By considering again a suitable subsequence, it is clear that $\left\{\frac{x^r-\overline{x}^r}{\|x^r-\overline{x}^r\|}\right\}_{r\in\mathbb{N}}$ may be assumed to converge to some $z\in\mathbb{R}^n$ with $\|z\|=1$. Hence letting $r\to\infty$ in (17) we obtain $u'_{t_i}z\le0$ for all $i=1,\ldots,n$. Consequently, (16) ensures

$$(u+\overline{c})'z\le0. \tag{18}$$

A completely symmetric argument entails $\overline{u}'_{\bar{t}_i}z\ge0$ for $i=1,\ldots,n$, and, hence,

$$(\overline{u}+\overline{c})'z\ge0. \tag{19}$$

This yields $u'z\le\overline{u}'z$. To show that we have even equality, we note that by convexity of f,

$$f(x^r)\ge f(\overline{x}^r)+(\overline{u}^r)'(x^r-\overline{x}^r)\quad\text{ and }\quad f(\overline{x}^r)\ge f(x^r)+(u^r)'(\overline{x}^r-x^r).$$

This implies

$$(\overline{u}^r)'(x^r-\bar{x}^r)\le f(x^r)-f(\overline{x}^r)\le(u^r)'(x^r-\bar{x}^r).$$

Hence, dividing by $\|x^r-\bar{x}^r\|$ and taking the limit yields $\overline{u}'z\le u'z$, which establishes $\overline{u}'z=u'z$. Consequently, expressions (18) and (19) coincide, and then

$$(u+\overline{c})'z=(\overline{u}+\overline{c})'z=0.$$

Finally, appealing to the first equality of (16), and recalling that $u'_{t_i}z\le0$ and $\lambda_i>0$ for all i, we conclude $u'_{t_i}z=0$ for $i=1,\ldots,n$. This, recalling that $z\neq0_n$, represents a contradiction with the fact that $\{u_{t_1},\ldots,u_{t_n}\}$ is a basis of $\mathbb{R}^n$. This completes the proof. □

Remark 11. Condition (10) is not necessary for metric regularity of the mapping $\mathcal{G}^*$. Just consider the optimization problem, in $\mathbb{R}^2$,

$$\begin{aligned}P(c,b):\quad&\text{Inf } x_1^2+x_2+c_1x_1+c_2x_2\\&\text{s.t. } x_1\ge b_1,\ x_2\ge b_2.\end{aligned}$$

Note that, in a neighborhood of $\left(\bar{c},\bar{b}\right)=(0_2,0_2)$, $\mathcal{F}^*$ is the Lipschitz function given by $\mathcal{F}^*(c,b)=\{(\max\{-c_1/2,b_1\},b_2)\}$, and then $\mathcal{G}^*$ is metrically regular at $\bar{x}=0_2$ for $\left(\bar{c},\bar{b}\right)$. However, condition (10) fails.

Remark 12. In fact, condition (10) is in general rather strong for metric regularity: as we will see, it implies a first order growth condition on f at $\overline{x}$ with respect to $\sigma(\bar{b})$, namely, the strong uniqueness of $\overline{x}$ as a minimizer of $P\left(\bar{c},\bar{b}\right)$ (see (20)), and moreover at least n constraints have to be active at $\overline{x}$. It is well known for finite nonlinear optimization problems with twice differentiable data that already certain second order growth conditions—which also typically hold in the situation of less than n active constraints—are sufficient and necessary for metric regularity of $\mathcal{G}^*$; see, e.g., [19, Chap. 8] and [20]. Generalizing this to the nonlinear semi-infinite case remains an open problem. However, in the next section we will see that for linear semi-infinite programs, condition (10) is indeed needed for metric regularity of $\mathcal{G}^*$ at $\overline{x}$ for $(\bar{c},\bar{b})$.

The rest of this section is concerned with the relationship between condition (10) and the strong uniqueness of a minimizer in the context of convex optimization. For continuously differentiable data f and g_t and under the Slater condition, property (ii) of Proposition 9 (recall that it is a consequence of condition (10)) is known as a sufficient condition for $\overline{x}$ to be a (locally) strongly unique minimizer of $P(\bar{c},\bar{b})$; see Theorem 3.1.16 in [14]. In the linear case, condition (10) turns out to be equivalent even to persistence of strong unicity under small parameter changes (see section 4 for details). In the following paragraphs we show how condition (10) is still sufficient for the latter property but no longer necessary.

Here, we say that $x \in \mathcal{F}(b)$ is a *strongly unique minimizer* of $P(c,b)$ if there exists a positive scalar α such that

$$f(y) + c'y \geq f(x) + c'x + \alpha \|y - x\| \text{ for all } y \in \mathcal{F}(b). \tag{20}$$

Obviously, in that case $\mathcal{F}^*(c,b) = \{x\}$. (Note that the convexity assumptions allow us to formulate the previous definition in global terms, not only in a neighborhood of x.) The following lemma characterizes the strong uniqueness of optimal solutions in terms of perturbations of vector c (which generalizes the linear version given in [9, Thm. 10.5]).

LEMMA 13. *A point x is the* strongly unique optimal solution *of $P(c,b)$ if and only if there exists $\varepsilon > 0$ such that $\|\widetilde{c} - c\| < \varepsilon$ implies $x \in \mathcal{F}^*(\widetilde{c},b)$ (in fact, for possibly smaller ε, x is the strongly unique solution of $P(\widetilde{c},b)$).*

Proof. According to [23, Chap. 5, Lem. 3] and [24, Thm. 23.8], x is a strongly unique optimal solution of $P(c,b)$ or, equivalently, of the problem

$$\operatorname*{Inf}_{x\in\mathbb{R}^n} \left\{ f(x) + c'x + \delta_{\mathcal{F}(b)}(x) \right\},$$

if and only if

$$0_n \in \operatorname{int}\{c + \partial(f + \delta_{\mathcal{F}(b)})(x)\} = c + \operatorname{int}\{\partial(f + \delta_{\mathcal{F}(b)})(x)\}$$

holds. The latter is equivalent to

$$0_n \in \widetilde{c} + \partial(f + \delta_{\mathcal{F}(b)})(x) \text{ for } \widetilde{c} \text{ close enough to } c,$$

i.e., $x \in \mathcal{F}^*(\widetilde{c},b)$ for $\widetilde{c}$ close enough to c. To ensure the last assertion, just take $\widetilde{c}$ such that $0_n \in \widetilde{c} + \operatorname{int}\left\{\partial(f + \delta_{\mathcal{F}(b)})(x)\right\}$. □

PROPOSITION 14. *If condition* (10) *holds at $\left(\overline{x},\left(\bar{c},\bar{b}\right)\right) \in gph(\mathcal{G}^*)$, then $\overline{x}$ is the strongly unique optimal solution of $P\left(\bar{c},\bar{b}\right)$.*

Proof. From Proposition 9(ii) there exist $u \in \partial f(\overline{x})$ as well as some $u_{t_i} \in -\partial g_{t_i}(\overline{x})$, $t_i \in T_{\bar{b}}(\overline{x})$, and some $\lambda_i > 0$ for $i \in \{1,\ldots,n\}$, such that $\{u_{t_1},\ldots,u_{t_n}\}$ is

a basis of $\mathbb{R}^n$ and

$$u + \bar{c} = \sum_{i=1}^{n} \lambda_i u_{t_i}.$$

So, $u + \bar{c} \in \text{int}(cone\,(\{u_{t_1}, \ldots, u_{t_n}\}))$. Hence, if $\|\widetilde{c} - \bar{c}\|$ is small enough, then

$$u + \widetilde{c} \in cone\,(\{u_{t_1}, \ldots, u_{t_n}\}),$$

which entails $\overline{x} \in \mathcal{F}^*\left(\widetilde{c}, \overline{b}\right)$. Thus, applying the previous lemma, $\overline{x}$ is the strongly unique optimal solution of $P\left(\overline{c}, \overline{b}\right)$. □

Remark 15. Actually, under condition (10), we have that $\left(\overline{c}, \overline{b}\right) \in \text{int}(\{(c, b) : P(c, b) \text{ has a strongly unique optimal solution}\})$ as a consequence of Proposition 9(i) and (v) (the latter ensures that all problems in a certain neighborhood have optimal solutions and (i) entails that these solutions are strongly unique). However, the converse statement does not hold. Just consider the parametrized convex problem, in which condition (10) fails trivially ($|T| = 1$, while the problem is posed in $\mathbb{R}^2$):

$$P(c, b) := \text{Inf}\{c_1 x_1 + c_2 x_2 \mid |x_1| - x_2 \leq b\},$$

around $\left(\overline{c}, \overline{b}\right) = ((0, 1)', 0)$. In fact, one can easily check that

$$\mathcal{F}^*(c, b) = \{(0, -b)\} \text{ if } \|c - \overline{c}\| < \frac{1}{\sqrt{2}},$$

and, since $\mathcal{F}^*(c, b)$ does not depend on c, we immediately conclude that $(0, -b)$ is a strongly unique optimal solution of $P(c, b)$ when $\|c - \overline{c}\| < \frac{1}{\sqrt{2}}$. (We used the Euclidean norm.)

Finally, note that the metric regularity property is sufficient neither for condition (9) nor for strong uniqueness. Just consider the example of Remark 11 and note that $\overline{x}$ is not a locally strongly unique minimizer of $P(0_2, 0_2)$, considering the feasible ray $\{(t, 0) \mid t \geq 0\}$.

4. Characterization of the metric regularity of $\mathcal{G}^*$ for linear problems. The following theorem provides the announced characterizations of the metric regularity of $\mathcal{G}^*$ for linear semi-infinite problems (4). Note that condition (v) is nothing else but (9). Moreover, condition (vi) comes from adapting to our notation Nürnberger's condition introduced in [22]. Actually, [22, Thm. 1.4] provides the counterpart of the equivalence (vi)⇔(vii) in the context in which perturbations of the a_t's are also allowed. The equivalence also holds, requiring only the boundedness of the a_t's, without continuity assumptions in the model (see [13, Thm. 4.1]).

THEOREM 16. *For the linear semi-infinite program* (4)*, let* $\left(\overline{x}, \left(\overline{c}, \overline{b}\right)\right) \in gph(\mathcal{G}^*)$. *Then the following conditions are equivalent:*

(i) $\mathcal{G}^*$ *is metrically regular at* $\overline{x}$ *for* $\left(\overline{c}, \overline{b}\right)$.

(ii) $\mathcal{F}^*$ *is strongly Lipschitz stable at* $\left(\left(\overline{c}, \overline{b}\right), \overline{x}\right)$.

(iii) $\mathcal{F}^*$ *is locally single-valued and continuous in some neighborhood of* $(\overline{c}, \overline{b})$.

(iv) $\mathcal{F}^*$ *is single-valued in some neighborhood of* $\left(\overline{c}, \overline{b}\right)$.

(v) $\sigma\left(\overline{b}\right)$ *satisfies the Slater condition and there is no* $D \subset T_{\overline{b}}(\overline{x})$ *with* $|D| < n$ *such that* $\overline{c} \in cone\,(\{a_t,\ t \in D\})$.

(vi) $\sigma\left(\overline{b}\right)$ *satisfies the Slater condition and for each* $D \subset T_{\overline{b}}(\overline{x})$ *with* $|D| = n$ *such that* $\overline{c} \in cone\,(\{a_t,\ t \in D\})$*; we have that all the possible subsets with* n *elements of* $\{a_t,\ t \in D\} \cup \{\overline{c}\}$ *are linearly independent.*

(vii) $\left(\bar{c},\bar{b}\right) \in \operatorname{int}\left(\{(c,b) : \mathcal{F}^*(c,b) \text{ consists of a strongly unique minimizer}\}\right)$.

Proof. The equivalence (i)⇔(ii) is nothing else but Lemma 5.

(ii)⇒(iii)⇒(iv). They are obvious consequences of the respective definitions.

(iv)⇒(v). From (iv) we immediately conclude that $\left(\bar{c},\bar{b}\right) \in \operatorname{int}\left(rge(\mathcal{G}^*)\right)$, which obviously implies $\bar{b} \in \operatorname{int}\left(rge(\mathcal{G})\right)$ and, from Lemma 3, $\mathcal{G}$ is metrically regular at $\overline{x}$ for $\bar{b}$ and $\sigma\left(\bar{b}\right)$ satisfies the Slater condition. In fact, if $S\left(\bar{b}\right)$ denotes the set of Slater points of $\sigma\left(\bar{b}\right)$, then one has $S\left(\bar{b}\right) = \operatorname{int}\left(\mathcal{F}\left(\bar{b}\right)\right)$ [9, Ex. 6.1]. Take $\widehat{x} \in S\left(\bar{b}\right)$ and define, for each $r \in \mathbb{N}$,

$$x^r := \overline{x} + \frac{1}{r}\left(\widehat{x} - \overline{x}\right) \in \operatorname{int}\left(\mathcal{F}\left(\bar{b}\right)\right)$$

(by the accessibility lemma).

Suppose, reasoning by contradiction, that $\bar{c} = \sum_{i=1}^{k} \lambda_i a_{t_i}$, with $t_i \in T_{\bar{b}}(\overline{x})$ and $\lambda_i > 0$ for $i = 1,\ldots,k$ and $k < n$. Now choose $u \in \{a_{t_1},\ldots,a_{t_k}\}^{\perp}$ with $\|u\| = 1$, whose existence is guaranteed by $k < n$. Then, since $x^r \in \operatorname{int}\left(\mathcal{F}\left(\bar{b}\right)\right)$, there exists some scalar α_r such that $y^r := x^r + \alpha_r u \in \mathcal{F}\left(\bar{b}\right)$, and we shall take $\alpha_r \in \left]0, 1/r\right]$. Define, for each $r \in \mathbb{N}$,

$$b_t^r := (1 - \varphi_r(t)) \min\{a_t' x^r, a_t' y^r\} + \varphi_r(t)\,\bar{b}_t,$$

where $\varphi_r : T \to [0,1]$ is a continuous function verifying

$$\varphi_r(t) = 0 \ \text{ if } t \in \{t_1,\ldots,t_k\} \quad \text{and} \quad \varphi_r(t) = 1 \ \text{ if } a_t'\overline{x} - \bar{b}_t \geq \tfrac{1}{r}.$$

The existence of such a φ_r is guaranteed by Urysohn's lemma. If $\left\{t \in T \mid a_t'\overline{x} - \bar{b}_t \geq \frac{1}{r}\right\}$ is empty, we take $\varphi_r \equiv 0$. Observe that $x^r, y^r \in \mathcal{F}\left(\bar{b}\right)$ implies that $x^r, y^r \in \mathcal{F}\left(b^r\right)$ for all r. Moreover, from the choice of u, we have $\{t_1,\ldots,t_k\} \subset T_{b^r}(x^r) \cap T_{b^r}(y^r)$, and $\bar{c} = \sum_{i=1}^{k} \lambda_i a_{t_i}$ ensures $x^r, y^r \in \mathcal{F}^*\left(\bar{c}, b^r\right)$ for all r (see Lemma 1). Now, let us show that $\lim_{r\to\infty} b^r = \bar{b}$. In fact, in the nontrivial case $a_t'\overline{x} - \bar{b}_t < \frac{1}{r}$ (otherwise $b_t^r = \bar{b}_t$) we have

$$\begin{aligned}
\left|b_t^r - \bar{b}_t\right| &\leq (1 - \varphi(t))\left|\min\{a_t' x^r, a_t' y^r\} - \bar{b}_t\right| \\
&\leq \max\left\{\left|a_t' x^r - \bar{b}_t\right|, \left|a_t' y^r - \bar{b}_t\right|\right\} \\
&\leq \left|a_t' x^r - \bar{b}_t\right| + \left|a_t'\left(y^r - x^r\right)\right| \\
&\leq \left|a_t'\left(x^r - \overline{x}\right)\right| + \left(a_t'\overline{x} - \bar{b}_t\right) + \|a_t\|_* \frac{1}{r} \\
&\leq \frac{1}{r}\left(1 + (1 + \|\widehat{x} - \overline{x}\|) \max_{t\in T} \|a_t\|_*\right),
\end{aligned}$$

just recalling the definition of x^r. Hence

$$\left\|b^r - \bar{b}\right\|_\infty \leq \frac{1}{r}\left(1 + (1 + \|\widehat{x} - \overline{x}\|) \max_{t\in T} \|a_t\|_*\right).$$

In this way, we provided a sequence $\{b^r\}_{r\in\mathbb{N}}$ converging to $\bar{b}$ such that $\mathcal{F}^*\left(\bar{c}, b^r\right)$ is not a singleton, which contradicts (iv).

(v)⇒(i). This follows from Theorem 10.

(v)⇔(vi) comes from standard arguments of linear algebra. Once we have established the equivalence among all conditions (i) to (vi), note that (vi)⇒(vii) comes from [22, Thm. 1.4] by taking into account that perturbations (c, b) are a particular case of perturbations of all coefficients. Finally, (vii)⇒(iv) is trivial. □

Remark 17. Example 4.6 in [20] shows that in the convex case (even for finite programs) the metric regularity of $\mathcal{G}^*$ (or, equivalently, strong Lipschitz stability of $\mathcal{F}^*$) does not necessarily hold if $\mathcal{F}^*$ is Kojima-stable (locally single-valued and continuous). This is in contrast to the linear semi-infinite case treated in the foregoing theorem.

5. Concluding remarks. The following diagram summarizes the main results of the paper concerning the convex case (1). The question of whether or not the strong uniqueness of an optimal solution for (c, b) near $(\bar{c}, \bar{b})$ implies the metric regularity of $\mathcal{G}^*$ at $(\overline{x}, (\bar{c}, \bar{b}))$ remains an open problem. Observe that condition (10) strictly implies the others in the diagram. Nevertheless, it is the only one which can be checked from the nominal problem's data, without involving parameters in a neighborhood.

$$\left\{\begin{array}{c}\text{Condition (10)}\\ \text{at } (\overline{x}, (\bar{c}, \bar{b})) \in gph\mathcal{G}^*\end{array}\right\} \underset{\substack{\Rightarrow\\ \not\Leftarrow}}{\overset{\text{Rem. 15}}{\underset{\text{Rem. 15}}{}}} \left(\bar{c}, \bar{b}\right) \in \operatorname{int}\left\{(c,b): \begin{array}{c}P(c,b) \text{ has a}\\ \text{strongly unique}\\ \text{minimizer}\end{array}\right\}$$

$$\text{Thm. 10} \Downarrow \quad \not\Uparrow \text{ Rem. 11} \qquad\qquad \not\Uparrow \text{ Rem. 15}$$

$$\left\{\begin{array}{c}\mathcal{G}^*\text{is metrically regular}\}\\ \text{at } (\overline{x}, (\bar{c}, \bar{b})) \in gph\mathcal{G}^*\end{array}\right\} \underset{\text{Lemma 5}}{\Leftrightarrow} \left\{\begin{array}{c}\mathcal{F}^* \text{ is single-valued and Lipschitz}\\ \text{in a neighborhood of } (\overline{x}, (\bar{c}, \bar{b}))\\ \text{(strongly Lipschitz stable)}\end{array}\right\}$$

$$\text{obvious} \Downarrow \quad \not\Uparrow \text{ Rem. 17}$$

$$\left\{\begin{array}{c}\mathcal{F}^* \text{ is single-valued and continuous}\\ \text{in a neighborhood of } (\overline{x}, (\bar{c}, \bar{b}))\\ \text{(Kojima-type stability)}\end{array}\right\}$$

When confined to the linear case, Theorem 16 establishes the equivalence among all of the conditions above.

Acknowledgment. The authors are indebted to the referees for their helpful critical comments.

REFERENCES

[1] B. Bank, J. Guddat, D. Klatte, B. Kummer, and K. Tammer, *Nonlinear Parametric Optimization*, Birkhäuser Verlag, Basel, Boston, 1983.
[2] C. Berge, *Topological Spaces*, Macmillan, New York, 1963.
[3] B. Brosowski, *Parametric semi-infinite linear programming* I. *Continuity of the feasible set and of the optimal value*, Math. Programming Stud., 21 (1984), pp. 18–42.
[4] M. J. Cánovas, A. L. Dontchev, M. A. López, and J. Parra, *Metric regularity of semi-infinite constraint systems*, Math. Program. Ser. B, 104 (2005), pp. 329–346.
[5] M. J. Cánovas, M. A. López, J. Parra, and M. I. Todorov, *Stability and well-posedness in linear semi-infinite programming*, SIAM J. Optim., 10 (1999), pp. 82–98.
[6] A. L. Dontchev, A. S. Lewis, and R. T. Rockafellar, *The radius of metric regularity*, Trans. Amer. Math. Soc., 355 (2003), pp. 493–517.
[7] T. Fischer, *Contributions to semi-infinite linear optimization*, in Approximation and Optimization in Mathematical Physics, B. Brosowski and E. Martensen, eds., Peter Lang, Frankfurt-Am-Main, Germany, 1983, pp. 175–199.

[8] V. E. Gayá, M. A. López, and V. N. Vera de Serio, *Stability in convex semi-infinite programming and rates of convergence of optimal solutions of discretized finite subproblems*, Optimization, 52 (2003), pp. 693–713.
[9] M. A. Goberna and M. A. López, *Linear Semi-Infinite Optimization*, John Wiley & Sons, Chichester, UK, 1998.
[10] M. A. Goberna, M. A. López, and M. I. Todorov, *Stability theory for linear inequality systems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 730–743.
[11] M. A. Goberna, M. A. López, and M. I. Todorov, *A generic result in linear semi-infinite optimization*, Appl. Math. Optim., 48 (2003), pp. 181–193.
[12] S. Gomez, A. Lancho, and M. Todorov, *Stability in convex semi-infinite optimization*, C. R. Acad. Bulgare Sci., 55 (2002), pp. 23–26.
[13] S. Helbig and M. I. Todorov, *Unicity results for general linear semi-infinite optimization problems using a new concept of active constraints*, Appl. Math. Optim., 38 (1998), pp. 21–43.
[14] R. Hettich and P. Zencke, *Numerische Methoden der Approximation und Semi-infinite Optimierung*, Teubner, Stuttgart, 1982.
[15] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms* I, Springer-Verlag, Berlin, 1993.
[16] W. W. Hogan, *Point-to-set maps in mathematical programming*, SIAM Rev., 15 (1973), pp. 591–603.
[17] D. Klatte, *Stability of stationary solutions in semi-infinite optimization via the reduction approach*, in Advances in Optimization, Lecture Notes in Econom. and Math. Systems 382, W. Oettli and D. Pallaschke, eds., Springer-Verlag, Berlin, 1992, pp. 155–170.
[18] D. Klatte and B. Kummer, *Stability properties of infima and optimal solutions of parametric optimization problems*, in Nondifferentiable Optimization: Motivations and Applications, V. F. Demyanov and D. Pallaschke, eds., Springer-Verlag, Berlin, 1985, pp. 215–229.
[19] D. Klatte and B. Kummer, *Nonsmooth Equations in Optimization: Regularity, Calculus, Methods and Applications*, Nonconvex Optim. Appl. 60, Kluwer Academic, Dordrecht, The Netherlands, 2002.
[20] D. Klatte and B. Kummer, *Strong Lipschitz stability of stationary solutions for nonlinear programs and variational inequalities*, SIAM J. Optim., 16 (2005), pp. 96–119.
[21] M. Kojima, *Strongly stable stationary solutions in nonlinear programs*, in Analysis and Computation of Fixed Points, S. M. Robinson, ed., Academic Press, New York, 1980, pp. 93–138.
[22] G. Nürnberger, *Unicity in semi-infinite optimization*, in Parametric Optimization and Approximation, B. Brosowski and F. Deutsch, eds., Birkhäuser, Basel, 1984, pp. 231–247.
[23] B. T. Polyak, *Introduction to Optimization*, Optimization Software, New York, 1987.
[24] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[25] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
[26] J.-J. Rückmann, *On existence and uniqueness of stationary points in semi-infinite optimization*, Math. Programming, 86 (1999), pp. 387–415.

SIAM J. OPTIM.
Vol. 18, No. 3, pp. 733–751

ON THE EXISTENCE OF SOLUTIONS TO DIFFERENTIAL INCLUSIONS WITH NONCONVEX RIGHT-HAND SIDES*

M. I. KRASTANOV†, N. K. RIBARSKA‡, AND TS. Y. TSACHEV§

Abstract. We study the existence of solutions of differential inclusions with upper semicontinuous right-hand sides. The investigation was prompted by the well-known Filippov examples. We define a new concept, "colliding on a set." In the case when the admissible velocities do not "collide" on the set of discontinuities of the right-hand side, we expect that at least one trajectory emanates from every point. If the velocities do "collide" on the set of discontinuities of the right-hand side, the existence of solutions is not guaranteed, as is seen from one of Filippov's examples. In this case we impose an additional condition in order to prove the existence of a solution starting at a point of the discontinuity set. For the right-hand sides under consideration, we assume the following: whenever the velocities "collide" on a set S there exist tangent velocities (belonging to the Clarke tangent cone to S) on a dense subset of S. Then we prove the existence of an ε-solution for every $\varepsilon > 0$. Under additional assumptions we can pass to the limit as $\varepsilon \to 0$ and obtain a solution of the considered differential inclusion.

Key words. differential inclusions with nonconvex right-hand sides, existence of solutions, colliding on a set

AMS subject classifications. 34A36, 34A60

DOI. 10.1137/060659077

1. Introduction.

We study the existence of solutions of the differential inclusion

$$\dot{x} \in F(x), \quad x(0) = x_0, \tag{1.1}$$

with upper semicontinuous right-hand side F. Our attention is focused on the existence in the autonomous case alone, and we are not going to discuss questions like uniqueness of the solution, continuous dependence on the initial conditions, etc.

Filippov proved the existence of a solution of (1.1) for the case of upper semicontinuous right-hand side F with *convex* and compact values (cf., for example, [14], [15], [16]). Since then the convexity assumption on the right-hand side has been of universal use in the calculus of variations and optimal control when differential inclusions are involved. Filippov's convexifying approach, however, because of its generality, does not always provide the best result. A longstanding open problem is the existence of solutions of differential inclusions with upper semicontinuous right-hand sides with nonconvex values. As is seen from the well-known Filippov examples (cf., for example, [16]), such a solution does not always exist.

The question of whether solutions exist for continuous F without convex values was raised by Hermes in [18] and solved by Filippov in [15]. This was generalized to the case of lower semicontinuous right-hand sides, by Bressan [5] (cf. also [4], [6], [7]) by means of the selection approach, and by Lojasiewicz in [20] and [21] by means of

*Received by the editors May 5, 2006; accepted for publication (in revised form) November 3, 2006; published electronically October 4, 2007.
http://www.siam.org/journals/siopt/18-3/65907.html

†Department of Biomathematics, Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Acad. G. Bonchev str., bl 8, 1113 Sofia, Bulgaria (krast@math.bas.bg).

‡Department of Mathematics and Informatics, University of Sofia, James Bourchier Boul. 5, 1126 Sofia, Bulgaria (ribarska@fmi.uni-sofia.bg).

§Department of Operations Research, Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Acad. G. Bonchev str., bl 8, 1113 Sofia, Bulgaria (tsachev@math.bas.bg).

Filippov's method. Existence results unifying the continuous and the convex cases were obtained by Olech in [22] (cf. also [19]) and by Lojasiewicz in [21], who proved existence assuming lower semicontinuity on an open set and convexity plus upper semicontinuity on its complement.

Bressan, Cellina, and Colombo studied in [9] the existence of solutions to differential inclusions with upper semicontinuous cyclically monotone right-hand sides. Recall that a multifunction A: $R^n \to R^n$ is called cyclically monotone if for every cyclical sequence $x_0, x_1, \ldots, x_N = x_0$ (N-arbitrary) and every sequence $y_i \in A(x_i)$, $i = 1, \ldots, N$, we have $\sum_{i=1}^{N} \langle x_i - x_{i-1}, y_i \rangle \geq 0$, where as usual $\langle ., . \rangle$ denotes the inner product of R^n. These multifunctions are exactly the upper semicontinuous ones, the graph of which is contained in the subdifferential of a proper convex function (cf. [25]). The last result along these lines, known to the authors, was obtained in [3] by Bounkhel and Haddad. A direct corollary of their result is that (1.1) has a solution if the right-hand side is an upper semicontinuous map with compact values, the graph of which is contained in the graph of the proximal subdifferential of a uniformly regular lower semicontinuous function. Another result, obtained by Veliov in [26], already proved to be very useful in studying invariance, stability, and attainability properties of a given compact set with respect to a differential inclusion. It yields the existence of solutions of differential inclusions with right-hand sides of the form

$$F(x) = \left\{ \eta \in G(x) : \ D^- \psi(x; \eta) \leq \phi(x) \right\},$$

where G is an upper semicontinuous map with compact convex values, ψ is a locally Lipschitz function, ϕ is an upper semicontinuous real-valued function, and D^- denotes the lower Dini derivative. Another interesting result was presented by Cellina and Ornelas in [10].

As we already mentioned, Bressan in [5] and Lojasiewicz in [20] and [21] proved the existence of a solution with lower semicontinuous right-hand side F, possibly not convex valued. On the other hand, having an arbitrary upper semicontinuous multivalued mapping F, it is well known (cf. Fort [17]) that it is not lower semicontinuous on a first Baire category set. But a set of first Baire categories may have a very complicated structure. However, if a positive real ε is fixed, then the set where F is not ε-lower semicontinuous is contained in a closed set with empty interior. (It is said that F is ε-lower semicontinuous at the point $\hat{x}$ if there exists an open set U containing $\hat{x}$ and such that $F(\hat{x}) \subset F(x) + \varepsilon B$ for each $x \in U$.) This is one of the possible motivations to work with ε-approximations of F.

Another motivation is given by a general fact about upper semicontinuous multivalued mappings, namely, a selection theorem of Srivatsa. To formulate it, we need the following definition (cf. Ribarska [23]).

DEFINITION 1.1. *Let X be a topological space and let*

$$\mathcal{U} = \{U_\alpha : \ 1 \leq \alpha < \alpha_0\}$$

be a well-ordered family of its subsets. It is said that $\mathcal{U}$ is a relatively open partitioning of X iff

(i) *U_α is contained in $X \setminus \left(\bigcup_{\beta<\alpha} U_\beta \right)$ and it is relatively open in it for every α;*
(ii) *$X = \bigcup_{1 \leq \alpha < \alpha_0} U_\alpha$.*

Note that each element of a relatively open partitioning is an intersection of an open and a closed set. Particular cases of such partitionings appear in [8] (cf. also [13]) in a proof (much simpler than the original one) for the lower semicontinuous

case, and in [1] in the definition of the so-called *patchy vector field* which was used for constructing piecewise constant stabilizing feedback control.

Now we formulate the above-mentioned selection theorem (Srivatsa [24]).

THEOREM 1.2. *Let $F : X \Rightarrow Y$ be an upper semicontinuous multivalued mapping, where X is a metric space and Y is a convex subset of a normed linear space. Then F has a selector f which is of the first Baire class and which is a uniform limit of functions, each of them constant on the elements of some relatively open partitioning of X.*

The simple structure of the functions uniformly approximating the selector f prompts us to study the problem of the existence of ε-solutions of the original problem, i.e., we are looking for an absolutely continuous function $x : [0,T] \to R^n$ such that

$$\begin{cases} \dot{x}(t) \in F(x(t)) + \varepsilon \bar{B} \ \text{ a.e. on } [0,T], \\ x(0) = x_0. \end{cases}$$

But the simple Filippov examples show that finding an ε-solution is not a trivial task. Consider

$$F_1(x) = \begin{cases} 1 & \text{if } x < 0, \\ \{-1,1\} & \text{if } x = 0, \\ -1 & \text{if } x > 0 \end{cases} \quad \text{and} \quad F_2(x) = \begin{cases} -1 & \text{if } x < 0, \\ \{-1,1\} & \text{if } x = 0, \\ 1 & \text{if } x > 0. \end{cases}$$

The inclusion $\dot{x} \in F_1(x)$ does not have a solution when $x(0) = 0$. It does not even have an ε-solution for $\varepsilon \in (0,1)$ and $x(0) = 0$. The inclusion $\dot{x} \in F_2(x)$ has a solution for any x_0.

The geometric intuition suggests that the reason for the nonexistence of a solution in the first Filippov example is that the corresponding vector field "collides" on the set of discontinuity of the right-hand side. In the present paper we define rigorously the concept of "colliding on a set" (section 3). In the case when the admissible velocities do not "collide" on the set of discontinuity of the right-hand side, one can expect that at least one trajectory emanates from every point. If the velocities do "collide" on the set of discontinuity of the right-hand side, the existence of solutions is not guaranteed, as is seen from the first Filippov example. In this case one needs to impose an additional condition in order to prove the existence of a solution starting at a point of the discontinuity set. For the right-hand sides under consideration, we assume the following: whenever the velocities "collide" on a set S there exist tangent velocities (belonging to the Clarke tangent cone to S) on a dense subset of S. Then we prove the existence of an ε-solution of (1.1) for every $\varepsilon > 0$ (section 4). Under additional assumptions we can pass to the limit as $\varepsilon \to 0$ and obtain a solution of (1.1) (section 5).

2. Preliminaries and technical lemmas. For any sets $S_1, S_2 \subseteq R^n$ and $\eta \in R$, we set $S_1 + \eta S_2 := \{s_1 + \eta s_2 : s_1 \in S_1, s_2 \in S_2\}$. Also, $B_r(p) := \{x \in R^n : \|x - p\| < r\}$, $\bar{B}_r(p)$ is the closure of $B_r(p)$ for all $p \in R^n$, $B := B_1(0)$, and $\bar{B} := \bar{B}_1(0)$.

The principal nonsmooth objects used in this paper are the proximal subgradient and normal cone, and here we review these concepts; see [11] for a complete treatment. Let $S \subseteq R^n$ be closed and $s \in S$. A vector $\zeta \in R^n$ is called a *proximal normal* vector of S at s provided there exists $\sigma = \sigma(\zeta, s) > 0$ so that

$$\langle \zeta, s' - s \rangle \le \sigma ||s' - s||^2 \quad \text{for all } s' \in S. \tag{2.1}$$

The set of all proximal normals of S at s is denoted by $\hat{N}_S(s)$ and is a convex cone. One can show (cf. [11, p. 25]) that for each $\delta > 0$ and $s \in S$, $\zeta \in \hat{N}_S(s)$ iff there exists $\sigma = \sigma(\zeta, s) > 0$ so that

$$\langle \zeta, s' - s \rangle \leq \sigma ||s' - s||^2 \quad \text{for all } s' \in S \cap \delta \mathcal{B}_n(s). \tag{2.2}$$

By $N_S(x)$ we denote the set of all limiting normals to S at x, which is defined as follows:

$$N_S(x) := \left\{ \zeta : \zeta = \lim_{n\to\infty} \zeta_n,\ \zeta_n \in \hat{N}_S(x_n),\ x = \lim_{n\to\infty} x_n \right\}.$$

The Bouligand tangent cone $T_S(x)$ to S at the point x is defined in the following way:

$$T_S(x) := \left\{ v : v = \lim_{n\to\infty} \frac{x_n - x}{t_n},\ x_n \in S,\ x = \lim_{n\to\infty} x_n,\ \lim_{n\to\infty} t_n = 0,\ t_n > 0 \right\}.$$

At last, we say that v belongs to the Clarke tangent cone $\hat{T}_S(x)$ to S at x if for each sequence $\{x_n\}_{n=1}^{\infty}$ of points of S and for each sequence $\{t_n\}_{n=1}^{\infty}$ of positive numbers decreasing to zero there exists a sequence $\{v_n\}_{n=1}^{\infty}$ converging to v such that $x_n + t_n v_n \in S$ for all n.

Recall that the *distance function* $d_S(\cdot) : R^n \to R$ is given by $d_S(x) := \min\{\|x - s\| : s \in S\}$. Then $\zeta \in \hat{N}_S(s)$ iff there exists $\varepsilon > 0$ such that $d_S(s + t\zeta) = t\|\zeta\|$ for all $t \in (0, \varepsilon)$.

For the related functional concept, assume that $f : R^n \to (-\infty, \infty]$ is lower semicontinuous and let $x \in \text{domain}(f) := \{x' : f(x') < \infty\}$. Then $\xi \in R^n$ is called a *proximal subgradient* for f at x provided there exist $\sigma > 0$ and $\eta > 0$ so that

$$f(x') \geq f(x) + \langle \xi, x' - x \rangle - \sigma \|x' - x\|^2 \quad \text{for all } x' \in B_\eta(x). \tag{2.3}$$

The set of all proximal subgradients for f at x is denoted by $\partial^P f(x)$. This set could be empty at some points, even for C^1 functions (e.g., the proximal subgradient of $f(x) = -|x|^{3/2}$ at $x = 0$ is empty). Recall that the *epigraph* $\text{epi } f \subseteq R^{n+1}$ of $f : R^n \to [-\infty, +\infty]$ is the set $\{(x, r) : r \geq f(x)\}$, and is closed iff f is lower semicontinuous. Two fundamental facts are as follows:

$$\xi \in \partial^P f(x) \Longleftrightarrow (\xi, -1) \in \hat{N}_{\text{epi} f}(x, f(x)) \quad \text{for all } x \in \text{domain}(f),$$

$$\zeta \in \partial^P d_S(x) \Longleftrightarrow x + d_S(x)\zeta \in \text{proj}_S(x) \quad \text{for all } x \in R^n \setminus S, \tag{2.4}$$

where $\text{proj}_S(x) := \{s \in S : d_S(x) = \|x - s\|\}$ is called the set of all *projections* of x into S.

LEMMA 2.1. *Let $S = U \cap G$, where U is an open subset of R^n and G is a closed subset of R^n. Let $F : U \Rightarrow R^n$ be an upper semicontinuous map with convex and compact values. Moreover, let the set $F(U)$ be bounded. Then there exists an open subset V of U containing S such that $\text{proj}_G(x) \subset U$ for each point $x \in V$. If*

$$\min \{\langle x - z, v \rangle : z \in \text{proj}_G(x),\ v \in F(z)\} \leq 0 \tag{2.5}$$

for every $x \in V$, then the set S is weakly invariant with respect to F, i.e., for each point $x \in S$ there exist $t_x > 0$ and an absolutely continuous function $\varphi : [0, t_x] \to S$ such that $\dot{\varphi}(t) \in F(\varphi(t))$ for almost every $t \in [0, t_x]$.

Proof. Define the set V to be

$$V := \left\{x \in R^n : d_G(x) < d_{R^n \setminus U}(x)\right\}.$$

Clearly, V is an open subset of U containing S. If $x \in V$, then for every point $z \in \mathrm{proj}_G(x)$ we have

$$\|z - x\| = d_G(x) < d_{R^n \setminus U}(x),$$

i.e., $z \notin R^n \setminus U$. Thus $\mathrm{proj}_G(x) \subset U$.

According to (2.5), for each $x \in V$ there exist $z_x \in \mathrm{proj}_G(x)$ and $v_x \in F(z_x)$ such that $\langle x - z_x, v_x\rangle \le 0$. For every $x \in V$, take any of these v_x and define $\theta(x) := v_x$.

Let x be an arbitrary point of S. Take $t_x > 0$ in such a way that $x + t_x.cB \subset V$, where c is an upper bound of the set $\{\|y\| : y \in F(x),\, x \in U\}$. Let

$$\pi := \{t_0 = 0, t_1, \ldots, t_{n-1}, t_n = t_x\}$$

be a partition of the interval $[0, t_x]$ and let φ_π be the corresponding Euler arc, i.e.,

$$\varphi_\pi(t_0) := x,\ \varphi_\pi(t) := \varphi_\pi(t_{i-1}) + (t - t_{i-1})\,\theta\left(\varphi_\pi(t_{i-1})\right)$$

for $t_{i-1} \le t \le t_i$ and $i = 1, 2, \ldots, n$.

According to the choice of t_x, $\varphi_\pi(\cdot)$ is well defined on $[0, t_x]$ and is contained in V. We have for every $i = 1, 2, \ldots, n$ and for every $t \in (t_{i-1}, t_i]$

$$\begin{aligned} d_S^2\left(\varphi_\pi(t)\right) &\le \left\|\varphi_\pi(t) - z_{\varphi_\pi(t_{i-1})}\right\|^2 = \left\|\varphi_\pi(t_{i-1}) + (t - t_{i-1})\,\theta\left(\varphi_\pi(t_{i-1})\right) - z_{\varphi_\pi(t_{i-1})}\right\|^2 \\ &\le \left\|\varphi_\pi(t_{i-1}) - z_{\varphi_\pi(t_{i-1})}\right\|^2 + (t - t_{i-1})^2\,\left\|\theta\left(\varphi_\pi(t_{i-1})\right)\right\|^2 \\ &\quad + 2(t - t_{i-1})\left\langle \varphi_\pi(t_{i-1}) - z_{\varphi_\pi(t_{i-1})}, \theta\left(\varphi_\pi(t_{i-1})\right)\right\rangle \\ &\le d_S^2\left(\varphi_\pi(t_{i-1})\right) + (t - t_{i-1})^2\, c^2 \le d_S^2\left(\varphi_\pi(t_{i-1})\right) + (t_i - t_{i-1})^2\, c^2. \end{aligned}$$

Denote

$$\mu_\pi := \max_{1 \le i \le n} (t_i - t_{i-1}).$$

Then summing up the above inequalities for $i = 1, 2, \ldots, k$ and for $t \in (t_{k-1}, t_k)$, we obtain

$$d_S^2\left(\varphi_\pi(t)\right) \le d_S^2\left(\varphi_\pi(t_0)\right) + c^2 \sum_{i=1}^{k} (t_i - t_{i-1})^2 \tag{2.6}$$

$$\le d_S^2\left(\varphi_\pi(t_0)\right) + c^2 \mu_\pi \sum_{i=1}^{k} (t_i - t_{i-1}) \le d_S^2\left(\varphi_\pi(t_0)\right) + c^2 \mu_\pi t_x = c^2 \mu_\pi t_x.$$

Now let $\{\pi_j\}_{j=1}^{\infty}$ be a sequence of partitions such that $\mu_{\pi_j} \to 0$ as $j \to \infty$ and necessarily $n_j \to \infty$. Then the family $\left\{\varphi_{\pi_j}\right\}_{j=1}^{\infty}$ is equicontinuous and uniformly bounded. According to the Arzelà–Ascoli theorem, some subsequence of $\left\{\varphi_{\pi_j}\right\}_{j=1}^{\infty}$ converges uniformly to a continuous function $\varphi(\cdot)$ on $[0, t_x]$. The limiting function

inherits the Lipschitz constant c on $[0, t_x]$, and hence it is absolutely continuous. Passing to the limits in (2.6) with $\pi = \pi_j$ as $j \to \infty$, we obtain

$$d_S^2(\varphi(t)) \leq 0$$

for every $t \in [0, t_x]$, i.e., for $t \in [0, t_x]$

$$\varphi(t) \in S. \tag{2.7}$$

Define the multivalued map

$$F_S(x) := \overline{co} \bigcup_{y \in \mathrm{proj}_S(x)} F(y) \quad \text{for } x \in V.$$

Because of the upper semicontinuity and the boundedness of $F(\cdot)$, and of the compactness of the set $\mathrm{proj}_G(x)$ for every $x \in V$, $F_S(x)$ is a compact and convex set. Since the multivalued map $x \Rightarrow \mathrm{proj}_G(x)$ is upper semicontinuous, the map $x \Rightarrow F_S(x)$ is also upper semicontinuous. Moreover, by construction $\theta(\cdot)$ is a selection of $F_S(\cdot)$. By Corollary 1.12 on p. 186 in [11], $\varphi(\cdot)$ is a solution of

$$\begin{cases} \dot{x}(t) \in F_S(x(t)) \ \text{ a.e. on } [0, t_x], \\ x(0) = x. \end{cases}$$

Because $F_S(\cdot)$ and $F(\cdot)$ coincide on the set S, (2.7) implies that $\varphi(\cdot)$ is a solution of

$$\begin{cases} \dot{x}(t) \in F(x(t)) \ \text{ a.e. on } [0, t_x], \\ x(0) = x. \end{cases}$$

This completes the proof. □

Remark. If in the assumptions of Lemma 2.1 we replace (2.5) by

$$\min \{\langle \zeta, v \rangle : v \in F(z)\} \leq 0 \tag{2.8}$$

for each $z \in S$ with $\hat{N}_S(z) \neq \emptyset$ and for each $\zeta \in \hat{N}_S(z)$, the conclusion of Lemma 2.1 remains true, because (2.8) implies (2.5).

LEMMA 2.2. *Let $G : X \Rightarrow R^n$ be an upper semicontinuous uniformly bounded map defined on a complete metric space X and let $\varepsilon > 0$ be fixed. Then there exists an open and dense subset W of X such that for each $x_0 \in W$ there exists an open neighborhood V of x_0 with*

$$G(x_0) \subset G(x) + \varepsilon B \tag{2.9}$$

whenever $x \in V$.

Proof. Let $M > 0$ be such that $\|G(x)\| \leq M$ for each $x \in X$. Take $\{y_1, y_2, \ldots, y_k\} \subset R^n$ to be a finite $\varepsilon/2$-net for the closed ball $M\bar{B}$ centered at the origin with radius M. The upper semicontinuity implies that the sets

$$X_i = \left\{ x \in X : G(x) \bigcap \left(y_i + \frac{\varepsilon}{2} \bar{B} \right) \neq \emptyset \right\}, \ i = 1, 2, \ldots, k,$$

are closed. We set

$$W := \bigcap_{i=1}^{k} \left((X \setminus X_i) \bigcup \text{ int } X_i \right).$$

The set W is an open and dense subset of X. Take an arbitrary $x_0 \in W$ and denote

$$I(x_0) = \left\{ i : G(x_0) \bigcap \left(y_i + \frac{\varepsilon}{2} \bar{B} \right) \neq \emptyset \right\}.$$

Now, we set

$$V := \{ x \in W : I(x) = I(x_0) \}.$$

Clearly, V is an open subset of X and $x_0 \in V$. To validate the assertion of Lemma 2.2, take $x \in V$. Then verify

$$G(x_0) \subset \bigcup_{i \in I(x_0)} \left(y_i + \frac{\varepsilon}{2} \bar{B} \right) = \bigcup_{i \in I(x)} \left(y_i + \frac{\varepsilon}{2} \bar{B} \right) \subset \bigcup_{y \in G(x)} \left(y + \varepsilon \bar{B} \right) = G(x) + \varepsilon \bar{B}.$$

This completes the proof. □

LEMMA 2.3. *Let $S = U \bigcap K$, where U is an open subset of R^n and K is a closed subset of R^n. Let $c > 0$ and $\varepsilon > 0$ be fixed. Then there exists a relatively open and dense in S subset W of S such that for each $x_0 \in W$ there exists an open neighborhood V of x_0 with*

$$\hat{T}_S(x) \bigcap c \, \bar{B} \subset T_S(x_0) \bigcap c \, \bar{B} + \varepsilon B \tag{2.10}$$

for every point $x \in V$.

Proof. Let S be endowed with the inherited topology of R^n and let $G : S \Rightarrow R^n$ be defined as

$$G(x) := N_S(x) \bigcap c \, \bar{B}.$$

According to Proposition 6.6 on p. 202 in [25], G is upper semicontinuous and uniformly bounded. Applying Lemma 2.2, we get an open relatively dense in S set W such that for every point $x_0 \in W$ there exists a relatively open neighborhood V of x_0 with

$$N_S(x_0) \bigcap c \, \bar{B} \subset N_S(x) \bigcap c \, \bar{B} + \varepsilon B \tag{2.11}$$

whenever $x \in V$.

Assume that (2.10) does not hold. Then there exists $w \in \hat{T}_S(y) \bigcap c \, \bar{B}$ with $y \in V$ such that

$$w \notin T_S(x_0) \bigcap c \, \bar{B} + \varepsilon B.$$

According to Proposition 6.27 on p. 219 in [25], there exists $v \in N_S(x_0) \bigcap B$ satisfying

$$d_{T_S(x_0)}(w) = \langle v, w \rangle.$$

Because of (2.11),

$$cv \in N_S(x) \bigcap c \, \bar{B} + \varepsilon B,$$

i.e., $cv = v_1 + \varepsilon v_2$ with $v_1 \in N_S(x) \bigcap c \, \bar{B}$ and $v_2 \in B$. Hence,

$$\langle v, w \rangle = \left\langle \frac{v_1}{c}, w \right\rangle + \frac{\varepsilon}{c} \langle v_2, w \rangle \leq 0 + \varepsilon = \varepsilon,$$

i.e., $d_{T_S(x_0)}(w) \leq \varepsilon$, which is a contradiction. Thus (2.10) holds true. □

3. Colliding on a set. Let D be an intersection of an open and a closed subset of R^n, let $x_0 \in D$ and $T > 0$, and let $F : D \Rightarrow R^n$ be an upper semicontinuous mapping with nonempty compact values which are uniformly bounded.

DEFINITION 3.1. *Let us fix an arbitrary $\varepsilon > 0$. It is said that F does not ε-collide from D to the point $\hat{x}$ of D iff there exist a subset A of D and a multivalued map $G : A \Rightarrow R^n$ defined on it such that*

(i) *the set A is an intersection of an open and a closed set, and $\hat{x} \in A \cap \mathrm{cl}(\mathrm{int}\ A)$ (here* int *A denotes the interior of A relatively in D);*

(ii) *G is an upper semicontinuous convex valued map with $G(x) \subset F(x) + \varepsilon B$ for each $x \in A$;*

(iii) *for each point $x \in (\partial A) \cap A$ (here (∂A) denotes the boundary of A relatively in D) and for each $\zeta \in \hat{N}_A(x)$ there exists $v \in G(x)$ with*

$$\langle \zeta,\, v \rangle \,\leq\, 0. \tag{3.1}$$

Lemma 2.1 shows that the differential inclusion $\dot{x} \in F(x)$ with $x(0) = x_0$ has an ε-solution on a small time interval whenever $x_0 \in A$ and D is open in R^n (here A and D relate to $F(\cdot)$ as in Definition 3.1).

PROPOSITION 3.2. *The following assertions hold true:*

(a) *If F is ε-lower semicontinuous at $\hat{x} \in D$, then F does not ε-collide from D to the point $\hat{x}$. Therefore the subset of D consisting of all points to which F does not ε-collide contains a dense relatively open subset of D.*

(b) *If F is upper semicontinuous and convex valued on an open neighborhood V of $\hat{x} \in D$, then F does not ε-collide from D to $\hat{x}$.*

Proof. Part (b) is obvious, because we can set A to be $V \cap D$ and G to be F. For part (a), let V be an open (in D) neighborhood of $\hat{x}$ with $F(\hat{x}) \subset F(x) + \varepsilon B$ whenever $x \in V$. We set $A := V$ and $G(x) := \hat{y}$ for every $x \in A$, $\hat{y}$ being an arbitrarily fixed point of $F(\hat{x})$. The last statement in part (a) is an immediate corollary of Lemma 2.2 and the first statement. □

Another class of multivalued maps which do not ε-collide at any point consists of the monotone maps and some of their generalizations (cf. [12]).

DEFINITION 3.3. *The operator $T : X \Rightarrow X^*$ is called hypomonotone if for every $x_0 \in X$ there exist $\delta > 0$ and $\rho > 0$ such that for all $x_1,\ x_2 \in B_\delta(x_0)$ and for all $\xi_i \in T(x_i)$, $i = 1, 2$, we have*

$$\langle \xi_2 - \xi_1, x_2 - x_1 \rangle \geq -\rho \|x_1 - x_1\|^2.$$

PROPOSITION 3.4. *Let D be an open subset of R^n and let $F : D \Rightarrow R^n$ be an upper semicontinuous hypomonotone multivalued map with nonempty compact values. Then F does not ε-collide from D to any point $\hat{x}$ of D for every $\varepsilon > 0$.*

Proof. Let us fix $\hat{x} \in D$ and $\varepsilon > 0$. Denote $R := \max\{\|y\| :\ y \in F(\hat{x})\}$ and choose $\hat{y}$ in $F(\hat{x})$ with $\|\hat{y}\| = R$. The Euclidean norm in R^n being uniformly convex, we are able to find $r > 0$ and $\alpha > 0$ such that the slice

$$S = \left\{ y \in \bar{B}_{R+r}(\theta) :\ \langle y, \hat{y} \rangle \geq R^2 - \alpha \right\}$$

has diameter less than ε. Let $\tilde{S}$ be the smaller slice

$$\tilde{S} = \left\{ y \in \bar{B}_{R+r}(\theta) :\ \langle y, \hat{y} \rangle \geq R^2 - \frac{\alpha}{2} \right\}.$$

The positive numbers $\rho > 0$ and $\delta > 0$ come from the definition of hypomonotonicity of F at $\hat{x}$. Put

$$\beta = \min\left\{\frac{\delta}{2}, \frac{\alpha}{4\rho R^2}\right\} > 0.$$

Then the set $C = (F^{-1}(\tilde{S}) \cap \bar{B}_\beta(\hat{x})) + \left[0, \frac{\alpha}{2\rho R^2}\right]\hat{y}$ is closed in D because of the upper semicontinuity of F at $\hat{x}$. Here $F^{-1}(\tilde{S}) = \{x \in D : \ F(x) \cap \tilde{S} \neq \emptyset\}$. The set $U = B_\beta(\hat{x}) \cap \{x \in D : \ F(x) \subset B_{R+r}(\theta)\}$ is relatively open in D and contains $\hat{x}$. Now $A = C \cap U$ and $G(x) := S$ for each $x \in A$ satisfy the conditions in Definition 3.1. Indeed, let $z \in A$, that is, $z = x + t\hat{y}$ and $z \in U$ with $\|x - \hat{x}\| \leq \beta$ and $F(x) \cap \tilde{S} \neq \emptyset$. Let z^* be an arbitrary element of $F(z)$ and $y \in F(x) \cap \tilde{S}$. Then $\langle z^* - y, z - x\rangle \geq -\rho\|z - x\|^2$; therefore for each $t \in \left(0, \frac{\alpha}{2\rho R^2}\right]$

$$\begin{aligned}\langle z^*, \hat{y}\rangle &= \frac{1}{t}\langle z^*, z - x\rangle \geq \frac{1}{t}\langle y, z - x\rangle - \frac{\rho}{t}\|z - x\|^2 \\ &= \langle y, \hat{y}\rangle - \rho t\|\hat{y}\|^2 \geq R^2 - \frac{\alpha}{2} - \rho R^2 t \geq R^2 - \alpha.\end{aligned}$$

Thus $F(z) \subset S$ whenever $t > 0$. Otherwise $z \in F^{-1}(\tilde{S}) \subset F^{-1}(S)$. We see that in both cases (ii) is satisfied for every $z \in A$ because diam $(S) < \varepsilon$ and so $S \subset F(z) + \varepsilon B$.

The above reasoning (applied with $x := \hat{x}$ and $y := \hat{y}$) shows that if $z = \hat{x} + t\hat{y}$ with $t \in (0, \beta)$, then $\langle z^*, \hat{y}\rangle > R^2 - \frac{\alpha}{2}$ for each $z^* \in F(z)$, that is, $F(z) \subset$ int $\tilde{S}$, and by the upper semicontinuity of F we find a (relatively) open set V with $z \in V$ and $V \subset F^{-1}(\text{int } \tilde{S}) \cap B_\beta(\tilde{x}) \subset C$. So $\hat{x} \in A \cap$ cl(int A) and (i) is satisfied. Now let $x \in \partial A$ and $\xi \in \hat{N}_A(x)$. Then $x \in A$ implies that $x = \tilde{x} + \tilde{t}\hat{y}$ with $t \in \left[0, \frac{\alpha}{2\rho R^2}\right)$ and $\tilde{x} \in F^{-1}(\tilde{S}) \cap B_\beta(\hat{x})$ ($\tilde{t} \neq \frac{\alpha}{2\rho R^2}$ because $\beta \leq \frac{\alpha}{4\rho R^2}$). Hence again the points $z = x + t\hat{y}$ belong to A for all $t > 0$ sufficiently small. This yields

$$\langle \xi, z - x\rangle \leq \sigma\|z - x\|^2, \text{ i.e., } \langle \xi, \hat{y}\rangle \leq \sigma t\|\hat{y}\|^2$$

for every sufficiently small $t > 0$. Hence $\langle \xi, \hat{y}\rangle \leq 0$ and $\hat{y} \in S = G(x)$. This completes the proof of the proposition. □

In particular, since every monotone map is hypomonotone, monotone (and, of course, cyclically monotone) maps do not ε-collide. Also, the proximal subdifferentials of uniformly regular lower semicontinuous functions (see Definition 3.5 below) are hypomonotone too.

DEFINITION 3.5 (cf. [3]). *Let U be a nonempty open subset of R^n and let $f : U \to R^n$ be a lower semicontinuous function. We say that f is uniformly regular over U if there exists $\beta \geq 0$ such that for all $x \in U$ and for all $\xi \in \partial^P f(x)$ one has*

$$\langle \xi, x' - x\rangle \leq f(x') - f(x) + \beta\|x' - x\|^2 \quad \textit{for all } x' \in U.$$

To prove that $\partial^P f(x)$ is hypomonotone, take arbitrary $x_1, x_2 \in U$ and $\xi_i \in \partial^P f(x_i)$, $i = 1, 2$. Then summing the inequalities

$$f(x_2) - f(x_1) - \langle \xi_1, x_2 - x_1\rangle \geq -\beta\|x_2 - x_1\|^2,$$

$$f(x_1) - f(x_2) - \langle \xi_2, x_1 - x_2\rangle \geq -\beta\|x_2 - x_1\|^2,$$

we obtain

$$\langle \xi_2 - \xi_1, x_2 - x_1\rangle \geq -2\beta\|x_2 - x_1\|^2.$$

Example 1. Let C be the graph of a function $f : R \to R$ which is not absolutely continuous on any subinterval of R. Consider $F : R^2 \Rightarrow R^2$ defined by

$$F(x_1,x_2)=\begin{cases}\{(1,1)\} & \text{if } x_2 < f(x_1);\\ \text{co}\,\{(1,-1),(1,1),(1,-2)\} & \text{if } x_2 = f(x_1);\\ \{(1,-1)\} & \text{if } x_2 \in \left[f(x_1)+\dfrac{1}{4k}, f(x_1)+\dfrac{1}{4k-1}\right];\\ \{(1,-2)\} & \text{if } x_2 \in \left[f(x_1)+\dfrac{1}{4k+2}, f(x_1)+\dfrac{1}{4k+1}\right];\\ \{(1,-1),(1,-2)\} & \text{if } x_2 \in \left(f(x_1)+\dfrac{1}{4k+3}, f(x_1)+\dfrac{1}{4k+2}\right);\\ \{(1,-1),(1,-2)\} & \text{if } x_2 \in \left(f(x_1)+\dfrac{1}{4k+1}, f(x_1)+\dfrac{1}{4k}\right),\end{cases}$$

where k denotes a positive integer. Then F is lower semicontinuous on $R^2 \setminus C$ and upper semicontinuous and convex valued on C. As C is a closed set, the differential inclusion $\dot{x} \in F(x)$ has a solution according to Theorem 1 of [21]. It is easy to see that F does not satisfy Definition 3.1 for every point of C and each $\varepsilon \in (0, 1/2)$. Nevertheless, there exists an upper semicontinuous convex-valued map $G : R^2 \Rightarrow R^2$ (not contained in the ε-neighborhood of F) such that every solution of $\dot{x} \in G(x)$ is a solution of $\dot{x} \in F(x)$.

Example 2. Consider $F : R^3 \Rightarrow R^3$ defined by

$$F(x_1,x_2,x_3)=\begin{cases}\{(0,0,-1)\} & \text{if } x_3 > 0,\\ \{(0,0,1)\} & \text{if } x_3 < 0,\\ \{(0,0,\pm 1),(0,1,0)\} & \text{if } x_3 = 0 \text{ and } x_2 < 0,\\ \{(0,0,\pm 1),(0,-1,0)\} & \text{if } x_3 = 0 \text{ and } x_2 > 0,\\ \{(0,0,\pm 1),(0,\pm 1,0),(-1,0,0)\} & \text{if } x_3 = x_2 = 0 \text{ and } x_1 > 0,\\ \{(0,0,\pm 1),(0,\pm 1,0),(1,0,0)\} & \text{if } x_3 = x_2 = 0 \text{ and } x_1 < 0,\\ \{(0,0,\pm 1),(0,\pm 1,0),(\pm 1,0,0),(0,0,0)\} & \text{if } x_3 = x_2 = x_1 = 0.\end{cases}$$

This map F is upper semicontinuous, and it does not satisfy Definition 3.1 for any $\varepsilon \in (0,1)$ with $D = R^3$ and $\hat{x} \in D_1 := \{(x_1,x_2,x_3) : \ x_3 = 0\}$; $D = D_1$ and $\hat{x} \in D_2 := \{(x_1,x_2,x_3) : \ x_2 = x_3 = 0\}$; $D = D_2$ and $\hat{x} = (0,0,0)$.

Example 3. Consider $F : R^2 \Rightarrow R^2$ defined by

$$F(x_1,x_2)=\begin{cases}\{(0,-1)\} & \text{if } x_1 > 0 \text{ and } x_2 > 0,\\ \{(0,1)\} & \text{if } x_1 > 0 \text{ and } x_2 < 0,\\ \left\{-\dfrac{(x_1,x_2)}{\sqrt{x_1^2+x_2^2}}\right\} & \text{if } x_1 < 0,\\ \{(-1,0)\} & \text{if } x_1 > 0 \text{ and } x_2 = 0,\\ \{(y_1,y_2) : \ y_1^2+y_2^2 = 1\} & \text{if } x_1 = x_2 = 0.\end{cases}$$

This map F is upper semicontinuous and does not satisfy Definition 3.1 for any $\varepsilon \in (0,1)$ with $D = R^2$ and $\hat{x} \in D_1 := \{(x_1,x_2) : \ x_1 \geq 0, \ x_2 = 0\}$. On the other

hand, this map satisfies Definition 3.1 with the set D_1 and all of its points. Evidently, there is no trajectory of $\dot{x} \in F(x) + \varepsilon\bar{B}$, $x(0) = (0,0)$. The reason for this is the fact that the map $x \Rightarrow N_{D_1}(x) \cap c\bar{B}$ is not ε-lower semicontinuous at the point $(0,0)$.

These examples motivate the next two definitions.

DEFINITION 3.6. *Let C be a relatively closed (in D) subset of D. It is said that F ε-collides from D to C iff C is the closure (in D) of the complement in D of the set*

$$\{x \in D : F \text{ does not } \varepsilon\text{-collide from } D \text{ to } x\} \cap \left\{x \in D : N_D(x) \bigcap c\bar{B} \text{ is } \varepsilon\text{-lower semicontinuous at } x\right\},$$

where $c := \sup\{\|y\| : y \in F(x), x \in D\}$.

DEFINITION 3.7. *Let C be a relatively closed (in D) subset of D. It is said that F collides to the set C iff there exist $\varepsilon > 0$, a decreasing well-ordered family $\{D_\alpha\}_{1 \le \alpha < \alpha_0}$, and a family of upper semicontinuous multivalued maps $F_\alpha : D_\alpha \to R^n$ with nonempty compact values such that*

(i) *$D_1 = D$, $F_1 \equiv F$, $C = D_{\alpha_0}$;*
(ii) *D_α is a nonempty closed subset of $D_{\alpha-1}$ if α is not a limit ordinal and $D_\alpha = \bigcap_{\beta<\alpha} D_\beta$ if α is a limit ordinal;*
(iii) *$F_\alpha(x) \subseteq \left(T_{D_\alpha}(x) + \varepsilon\bar{B}\right) \bigcap F(x)$ for every $x \in D_\alpha \setminus D_{\alpha+1}$;*
(iv) *F_α ε-collides from D_α to $D_{\alpha+1}$.*

4. Basic assumption and existence of ε-solutions. The main idea of our approach is to assume that whenever the admissible velocities do "collide" on some set S, in order to have a solution of (1.1) which does not leave S, there exist tangent velocities to S. It is natural to assume that these velocities belong to the Bouligand tangent cone to S. The following example shows that this cone is too big for our purposes. The reason is that it lacks good continuity properties.

Example 4. We define the multivalued map $F : [0,1] \Rightarrow \{-1,1\}$ using a Cantor-like construction:

$$F(x) := \begin{cases} \{1\} & \text{if } x \in \left(\frac{1}{5}, \frac{2}{5}\right), \\ \{-1\} & \text{if } x \in \left(\frac{3}{5}, \frac{4}{5}\right), \\ \{1\} & \text{if } x \in \left(\frac{1}{5^2}, \frac{2}{5^2}\right) \cup \left(\frac{2}{5} + \frac{3}{5^2}, \frac{2}{5} + \frac{4}{5^2}\right) \cup \left(\frac{4}{5} + \frac{1}{5^2}, \frac{4}{5} + \frac{2}{5^2}\right), \\ \{-1\} & \text{if } x \in \left(\frac{3}{5^2}, \frac{4}{5^2}\right) \cup \left(\frac{2}{5} + \frac{1}{5^2}, \frac{2}{5} + \frac{2}{5^2}\right) \cup \left(\frac{4}{5} + \frac{3}{5^2}, \frac{4}{5} + \frac{4}{5^2}\right), \\ \ldots\ldots & \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{cases}$$

We define $F(x)$ to be $\{-1,1\}$ on the remaining Cantor set A. Clearly, A is the set of all points x at which F is not continuous. It is easy to check that F collides to A and $F(x) \bigcap T_A(x) \neq \emptyset$ for all $x \in A$. On the other hand, the differential inclusion $\dot{x} \in F(x)$ has no solution whenever the starting point belongs to a dense subset of A. Moreover, this differential inclusion has no ε-solution for each $\varepsilon \in (0,1)$ starting from the same points.

Let D be an intersection of an open and a closed subset of R^n, and let $F : D \Rightarrow R^n$ be an upper semicontinuous mapping with nonempty compact values which are uniformly bounded. Example 4 and Lemma 2.3 motivate the following assumption.

Basic assumption. For every $C \subseteq D$ such that F collides to C there exists a dense subset E of C such that

$$F(x) \bigcap \hat{T}_C(x) \neq \emptyset \quad \text{for all } x \in E.$$

It is straightforward to verify that Examples 1, 3, and 4 do not satisfy the basic assumption, and Example 2 satisfies it.

THEOREM 4.1. *Let D be an open subset of R^n and let $F : D \Rightarrow R^n$ be an upper semicontinuous mapping with nonempty compact values which are uniformly bounded. Moreover, let F satisfy the basic assumption. Then for each $\varepsilon > 0$, for each $T > 0$, and for each point x_0 of D there exists an absolutely continuous function $x : [0, T] \to R^n$ such that*

$$\begin{cases} \dot{x}(t) \in F(x(t)) + \varepsilon \overline{B} \quad \textit{a.e. on } [0, T], \\ x(0) = x_0. \end{cases}$$

Proof. We set $c := \sup\{\|y\| : \ y \in F(x), x \in D\}$ and $D_1 := D$. Let D_2 be the closed subset of D_1 to which F ε-collides from D_1. We set $W_1 := D_1 \setminus D_2$ and $F_1 \equiv F$. Then Proposition 3.2(a) implies that W_1 is a dense and relatively open (in D_1) subset of D_1.

According to the basic assumption, there exists a dense subset E_2 of D_2 such that

$$F(x) \bigcap \hat{T}_{D_2}(x) \neq \emptyset \quad \text{for all } x \in E_2. \tag{4.1}$$

We define $F_2 : D_2 \Rightarrow R^n$ as follows:

$$F_2(x) = \left\{ v \in R^n : \ x_n \to x, x_n \in E_2, \ v_n \to v, v_n \in F_1(x_n) \cap \hat{T}_{D_2}(x_n) \right\}.$$

It is straightforward to check that F_2 is an upper semicontinuous mapping with nonempty and compact values. Because of the upper semicontinuity of F, we have $F_2(x) \subseteq F(x)$ for each $x \in D_2$.

According to Lemma 2.3, there exists a relatively open and dense (in D_2) set $\tilde{W}_2$ such that for each point $\hat{x} \in \tilde{W}_2$ there exists an open neighborhood V of $\hat{x}$ with

$$\hat{T}_{D_2}(x) \bigcap c\,\bar{B} \subset T_{D_2}(\hat{x}) \bigcap c\,\bar{B} + \varepsilon \bar{B} \tag{4.2}$$

for every point $x \in V$. Let $v \in F_2(\hat{x})$ be arbitrary. Then there exist a sequence $\{x_n\}_{n=1}^{\infty}$ of points in E_2 converging to $\hat{x}$ and a sequence of velocities $v_n \to v$ with $v_n \in F_1(x_n) \bigcap \hat{T}_{D_2}(x_n)$. Then the upper semicontinuity of F_2 and (4.2) imply that

$$F_2(\hat{x}) \subseteq T_{D_2}(\hat{x}) \bigcap c\,\bar{B} + \varepsilon \bar{B}. \tag{4.3}$$

Let D_β and F_β be defined for each $\beta < \alpha$. If α is a limit ordinal, we put

$$D_\alpha = \bigcap_{\beta < \alpha} D_\beta, \ F_\alpha := F.$$

If α is not a limit ordinal, there exists $\beta = \alpha - 1 < \alpha$. Let D_α be the closed subset of $D_{\alpha-1}$ to which $F_{\alpha-1}$ ε-collides from $D_{\alpha-1}$.

We repeat the construction of F_2: According to the basic assumption, there exists a dense subset E_α of D_α such that

$$F(x) \bigcap \hat{T}_{D_\alpha}(x) \neq \emptyset \quad \text{for all } x \in E_\alpha. \tag{4.4}$$

We define $F_\alpha : D_\alpha \Rightarrow R^n$ as follows:

$$F_\alpha(x) = \left\{ v \in R^n : x_n \to x, x_n \in E_\alpha,\, v_n \to v, v_n \in F(x_n) \cap \hat{T}_{D_\alpha}(x_n) \right\}.$$

It is straightforward to check that F_α is an upper semicontinuous mapping with nonempty and compact values. Because of the upper semicontinuity of F, we have $F_\alpha(x) \subseteq F(x)$ for each $x \in D_\alpha$.

According to Lemma 2.3, there exists a relatively open and dense (in D_α) set $\tilde{W}_\alpha$ such that for each point $\hat{x} \in \tilde{W}_\alpha$ there exists an open neighborhood V of $\hat{x}$ with

$$\hat{T}_{D_\alpha}(x) \bigcap c\,\bar{B} \subset T_{D_\alpha}(\hat{x}) \bigcap c\,\bar{B} + \varepsilon \bar{B} \tag{4.5}$$

for every point $x \in V$. Let $v \in F_\alpha(\hat{x})$ be arbitrary. Then there exist a sequence $\{x_n\}_{n=1}^\infty$ of points in E_α converging to $\hat{x}$ and a sequence of velocities $v_n \to v$ with $v_n \in F(x_n) \bigcap \hat{T}_{D_\alpha}(x_n)$. Then the upper semicontinuity of F_α and (4.2) imply that

$$F_\alpha(\hat{x}) \subseteq T_{D_\alpha}(\hat{x}) \bigcap c\,\bar{B} + \varepsilon \bar{B}. \tag{4.6}$$

Thus we have constructed D_α and F_α. Let α_0 be the first ordinal number with $D_{\alpha_0} = \emptyset$.

Let x be an arbitrary point of D_1. We set

$$\alpha_x := \min \{\alpha : x \notin D_\alpha\}.$$

Let us assume that α_x is a limit ordinal, i.e., $\alpha_x = \sup\{\beta : \beta < \alpha_x\}$. The definition of α_x implies that for each $\beta < \alpha_x$ the point x belongs to D_β. Hence, $x \in \bigcap_{\beta<\alpha_x} D_\beta = D_{\alpha_x}$. The last inclusion contradicts the definition of α_x. So, α_x is not a limit ordinal. Then $x \in D_{\alpha_x - 1} \setminus D_{\alpha_x}$, and hence $F_{\alpha_x - 1}$ does not ε-collide to x from $D_{\alpha_x - 1}$. According to Definition 3.1, there exists a subset A of $D_{\alpha_x - 1}$ and a multivalued map $G : A \Rightarrow R^n$ such that

(i) the subset A is an intersection of an open and a closed set, and the point x belongs to $A \bigcap \mathrm{cl}\,(\mathrm{int}\, A)$;

(ii) G is an upper semicontinuous convex-valued map with $G(\hat{x}) \subset F_{\alpha_x - 1}(\hat{x}) + \varepsilon \bar{B}$ for each $\hat{x} \in A$;

(iii) for each point $\hat{x} \in \partial A \bigcap A$ and for each $\zeta \in \hat{N}_A(\hat{x})$ there exists $v \in G(\hat{x})$ with

$$\langle \zeta,\, v \rangle \leq 0. \tag{4.7}$$

Let $\hat{x}$ be an arbitrary point of A and let ζ be an arbitrary element of $\hat{N}_A(\hat{x})$. If $\hat{x}$ is in the interior of A relatively in $D_{\alpha_x - 1}$, then $\zeta \in \hat{N}_{D_{\alpha_x - 1}}(\hat{x})$. According to (4.6) and (ii),

$$G(\hat{x}) \subset F_{\alpha_x - 1}(\hat{x}) + \varepsilon \bar{B} \subset T_{D_{\alpha_x - 1}}(\hat{x}) \cap c\bar{B} + 2\varepsilon \bar{B},$$

and so there exists a vector $v \in T_{D_{\alpha_x - 1}}(\hat{x}) \cap (G(\hat{x}) + 2\varepsilon \bar{B})$. Hence, $\langle \zeta,\, v \rangle \leq 0$ holds true. If $\hat{x}$ is in the boundary of A relatively in $D_{\alpha_x - 1}$, then by (ii) and (iii) there exists $v \in G(\hat{x}) \subset F_{\alpha_x - 1}(\hat{x}) + \varepsilon \bar{B}$ satisfying (4.7). So, we can apply Lemma 2.1 to

the set A and the map $G + 2\varepsilon\bar{B}$. Thus we obtain the existence of $t_x > 0$ and of a trajectory $\varphi : [0, t_x) \to A$ which is a solution of

$$\begin{cases} \dot{x}(t) \in G(x(t)) + 2\varepsilon\bar{B} \subseteq F_{\alpha_x - 1}(x(t)) + 3\varepsilon\bar{B} \ \text{ a.e. on } [0, t_x], \\ x(0) = x. \end{cases}$$

In particular, there exist $t_{x_0} > 0$ and an absolutely continuous function $\varphi : [0, t_{x_0}) \to D_1$ which is well defined on $[0, t_x)$ and is a solution of the following differential inclusion:

$$\begin{cases} \dot{x}(t) \in F(x(t)) + 3\varepsilon\,\bar{B} \ \text{ a.e. on } [0, t_{x_0}), \\ x(0) = x_0. \end{cases}$$

Let $\hat{T}$ be the supremum of all $\bar{T} \geq 0$ for which the function φ has an absolutely continuous extension on the interval $[0, \bar{T})$ (which we denote again by φ) which is a solution of the above-written differential inclusion on the interval $[0, \bar{T})$. Clearly, $\hat{T} \geq t_{x_0} > 0$. Let us assume that $\hat{T} \leq T$. Then we have that

$$\begin{cases} \dot{\varphi}(t) \in F(\varphi(t)) + 3\varepsilon\,\bar{B} \ \text{ a.e. on } [0, \hat{T}), \\ \varphi(0) = x_0. \end{cases} \tag{4.8}$$

Then for each increasing sequence $\{t_n\}_{n=1}^{\infty} \to \hat{T}$ we have

$$\|\varphi(t_n) - \varphi(t_{n-1})\| \leq \int_{t_{n-1}}^{t_n} \|\dot{\varphi}(\tau)\|\, d\tau \leq c\,(t_n - t_{n-1}).$$

This means that the sequence $\{\varphi(t_n)\}_{n=1}^{\infty}$ is a Cauchy sequence. Hence, there exist $\varphi(\hat{T}) := \lim_{t \to \hat{T}} \varphi(t)$. But then we can find an absolutely continuous extension of the function φ which is a solution of (4.8) on the interval $\left[0, \hat{T} + t_{\varphi(\hat{T})}\right)$. This contradicts the definition of $\hat{T}$. Hence, $\hat{T} > T$ and this completes the proof. □

COROLLARY 4.2. *Let S be a closed set and let F be a uniformly bounded upper semicontinuous multivalued map with nonempty compact values which satisfies the basic assumption. If*

$$\max_{\xi \in \hat{N}_S(x)} \min_{v \in F(x)} \langle \xi, v \rangle \leq 0$$

for every $x \in S$, then (S, F_ε) is viable (weakly invariant) for each $\varepsilon > 0$, i.e., for each point $x_0 \in S$ there exists a solution $x(\cdot)$ of the differential inclusion $\dot{x} \in F_\varepsilon(x)$, $x(0) = x_0$, with $F_\varepsilon(x) = F(x) + \varepsilon\bar{B}$, defined on some interval $[0, t_{x_0})$, $t_{x_0} > 0$, such that $x(t) \in S$ for each $t \in [0, t_{x_0})$.

In contrast to the convex-valued case (cf., for example, [2] or [11]), the basic assumption is crucial for the validity of this corollary.

5. Existence of a solution. The question of when it is possible to pass to the limit as $\varepsilon \to 0$ and to obtain a real solution of (1.1) remains open. Assuming that F is an upper semicontinuous mapping satisfying the basic assumption, a relatively open partitioning $\mathcal{U}_\varepsilon$ is constructed in the proof of Theorem 4.1 for every $\varepsilon > 0$. Its elements are invariant with respect to the trajectories of the differential inclusion $\dot{x} \in F(x) + \varepsilon\bar{B}$.

The next theorem is inspired by Bressan's proof in the lower semicontinuous case. To prove it, we have additionally assumed that some refinements of the relatively open partitioning $\mathcal{U}_\varepsilon$ may be chosen finite and that their elements are invariant with respect to the trajectories of the differential inclusion $\dot{x} \in F(x) + \eta\bar{B}$ for each $0 < \eta < \varepsilon$.

THEOREM 5.1. *Let the set D be an intersection of an open and a closed subset of R^n and let $F : D \Rightarrow R^n$ be a uniformly bounded upper semicontinuous multivalued mapping with nonempty compact values. Let*

$$\mathcal{U}^k = \left\{U_\alpha^k : \ 1 \le \alpha < \alpha_0^k\right\}, \quad k = 1, 2, \dots,$$

be uniformly locally finite (i.e., every point of D has a neighborhood V which intersects at most finitely many members of $\mathcal{U}^k$ for each positive integer k) relatively open partitionings of D. Let $\mathcal{U}^{k+1}$ be a refinement of $\mathcal{U}^k$ for each k. Moreover, we assume that there exist multivalued mappings $F_k : D \Rightarrow R^n$ such that the restriction of F_k on an arbitrary element U of $\mathcal{U}^k$ is upper semicontinuous with nonempty convex compact values, the diameter of $F_k(U)$ is less than or equal to $\frac{1}{k}$, and $F_{k+1} \subset F_k$ and $F_k \subset F + \frac{1}{k}\bar{B}$ for each k. We also assume that for each point x_0 of D there exists $T > 0$ such that

(i) *for each positive integer k there exists an absolutely continuous function $x_k : [0, T] \to D$ such that*

$$\begin{cases} \dot{x}(t) \in F_k(x(t)) \ \ \textit{a.e. on } [0, T], \\ x(0) = x_0; \end{cases} \tag{5.1}$$

(ii) *for each positive integer k and m with $k \ge m$, and for each solution $x_k(\cdot)$ of (5.1), the function*

$$t \to \alpha(t), \ \textit{where } x_k(t) \in U_{\alpha(t)}^m,$$

is monotone increasing.

Then there exists an absolutely continuous function $x : [0, T] \to D$ such that

$$\begin{cases} \dot{x}(t) \in F(x(t)) \ \ \textit{a.e. on } [0, T], \\ x(0) = x_0. \end{cases}$$

Proof. Let us fix an arbitrary point $x_0 \in D$. Without loss of generality, we may think that $\{x_k(t) : \ t \in [0, T]\}$ is contained in the neighborhood V of x_0 that intersects at most finitely many members of $\mathcal{U}^k$ for every positive integer k. Let $\alpha_1^k < \alpha_2^k < \dots < \alpha_{s(k)}^k$ be such that $V \cap U_\alpha^k = \emptyset$ whenever $\alpha \ne \alpha_s^k$, $s = 1, \dots, s(k)$.

Let us fix for a moment the positive integer m and let k be an arbitrary integer with $k \ge m$. Then $\left\{x_k(t) : \ t \in [0, T]\right\} \subset \bigcup_{s=1}^{s(m)} U_{\alpha_s^m}^m$. We set

$$\Delta_s^{m,k} := \left\{t : \ x_k(t) \in U_{\alpha_s^m}^m\right\}, \quad s = 1, 2, \dots, s(m).$$

According to (ii), the sets $\Delta_s^{m,k}$, $s = 1, 2, \dots, s(m)$, are intervals such that $t_{s_1} < t_{s_2}$ for each point $t_{s_i} \in \Delta_{s_i}^{m,k}$, $i = 1, 2$, whenever $s_1 < s_2$. Of course, some of the sets $\Delta_s^{m,k}$ can be empty.

Without loss of generality (cf., for example, [21]) we may think that the sequence $\{x_k(\cdot)\}_{k=1}^\infty$ is uniformly convergent on the interval $[0, T]$ to an absolutely continuous function denoted by $x(\cdot)$.

Let us denote by $t_s^{m,k}$, $s = 1, 2, \ldots, s(m)$, the points of transition of the trajectory $x_k(\cdot)$ from the set $U^m_{\alpha^m_{s-1}}$ to the set $U^m_{\alpha^m_s}$. If the trajectory passes from $U^m_{\alpha^m_s}$ to $U^m_{\alpha^m_l}$ with $l \geq s+2$, then we set $t_s^{m,k} = t_{s+1}^{m,k} = \cdots = t_{l-1}^{m,k} = t_l^{m,k}$. Without loss of generality, we may think that each sequence $\left\{t_s^{m,k}\right\}_{k=1}^{\infty}$ is monotone and tends to the number t_s^m, $s = 1, 2, \ldots, s(m)$. The inequalities $t_s^{m,k} \leq t_{s+1}^{m,k}$ imply that $t_s^m \leq t_{s+1}^m$, $s = 1, 2, \ldots, s(m)$.

For almost all τ from $[0, T]$ (cf. Lemma 2 from [21]) the derivatives $\dot{x}(\tau)$ and $\dot{x}_k(\tau)$ exist and

$$\dot{x}(\tau) \in \ \overline{\text{co}}\ \{\dot{x}_k(\tau) : \ k \geq k_0\}$$

for any positive integer k_0.

If τ does not coincide with any of the points t_s^m, $s = 1, 2, \ldots, s(m)$, then $\tau \in (t_{s-1}^m, t_s^m)$ for some s, and so $x_k(\tau) \in U^m_{\alpha^m_{s-1}}$ for all sufficiently large k. If τ coincides with the point t_s^m for some s, then two cases are possible: the sequence $\left\{t_s^{m,k}\right\}_{k=1}^{\infty}$ is nonincreasing, and hence $x_k(\tau) \in U^m_{\alpha^m_s}$ for all sufficiently large k, or the sequence $\left\{t_s^{m,k}\right\}_{k=1}^{\infty}$ is increasing, and therefore $x_k(\tau) \in U^m_{\alpha^m_{s-1}}$ for all sufficiently large k.

In all cases $x_k(\tau)$, $k \geq k_0$, belong to one and the same element of the partitioning $\mathcal{U}^m$, say U, and thus

$$\dot{x}(\tau) \in \ \overline{\text{co}}\ \{\dot{x}_s(\tau) : \ s \geq k_0\} \subset \ \overline{\text{co}}\ \left(\bigcup_{s \geq k_0} F_s(x_s(\tau))\right) \subset \overline{\text{co}}\ \left(\bigcup_{s \geq k_0} F_m(x_s(\tau))\right)$$

$$\subset \ \overline{\text{co}}\ \left(F_m(x_k(\tau)) + \frac{1}{m}\bar{B}\right) = F_m(x_k(\tau)) + \frac{1}{m}\bar{B} \subset F(x_k(\tau)) + \frac{2}{m}\bar{B}$$

for all $k \geq \max(m, k_0)$. Now the upper semicontinuity of F and $x_k(\tau) \to x(\tau)$ imply

$$\dot{x}(\tau) \in \ F(x(\tau)) + \frac{2}{m}\bar{B}. \tag{5.2}$$

Since $\dot{x}(\tau)$ does not depend on m and F is compact valued, we obtain that

$$\dot{x}(\tau) \in \ F(x(\tau)).$$

This completes the proof. □

Remark 1. It is possible to consider the lower semicontinuous case as a corollary to this result. Indeed, let us consider the simplest case when D is a closed set. $G : D \to R^n$ is a uniformly bounded lower semicontinuous multivalued map with compact values such that $G(y) \subset T_D(y)$ for each $y \in D$. We set $x = (y, z)$, $F(x) := G(y) \times \{1\}$ and consider the differential inclusion $\dot{x} \in F(x)$. Starting from the relatively open partitioning constructed in the proof of Lemma 6.2 in [13, p. 67], one can refine it by dividing if necessary the elements of the partitioning by the finitely many hyperplanes $\{(y, z) : \ z = z_i\}$, where z_i is the sum of δ_i and the last coordinate of ω_i (these notations are from the proof of Lemma 6.2 in [13, p. 67]). Then the obtained partitioning can be ordered in such a way as to satisfy the assumptions of Theorem 5.1 with $F_k(x) = \left(f_k(y) + \frac{1}{2^{k-1}}\bar{B}\right) \times \{1\}$ (again f_k is from the above-mentioned proof). Note that in the proof of Theorem 5.1 the upper semicontinuity of F is used only to derive (5.2). In the lower semicontinuous case the limit trajectory $x(\cdot)$ cannot remain on the boundaries of the elements of the partitioning. Hence, for all τ except for finitely

many values, $x(\tau)$ belongs to the interior of the elements of the partitioning and we can pass to the limit as $k \to \infty$ without using the upper semicontinuity.

Remark 2. Theorem 5.1 can be applied to the differential inclusions with right-hand sides F_2 (cf. the end of the introduction), the multivalued mappings described in Example 2, Example 3 (if $(0,0)$ is added to $F(0,0)$), and Example 4 (if the origin is added to $F(x)$ for every $x \in A$). These multivalued mappings, clearly, are not lower semicontinuous everywhere. Moreover, the assumptions of Theorem 5.1 do not exclude the existence of periodic solutions (cf. the assumption (ii)). To show this, we consider the following example: We set $D = \bigcup_{i=1}^{4} D_i$, where

$$
\begin{aligned}
D_1 &= \{(x,y):\ x>0,\ y\geq 0,\ x+y<1\},\\
D_3 &= \{(x,y):\ x<0,\ y\leq 0,\ x+y>-1\},\\
D_2 &= \{(x,y):\ x\leq 0,\ y>0,\ y-x<1\},\\
D_4 &= \{(x,y):\ x\geq 0,\ y<0,\ y-x>-1\}.
\end{aligned}
$$

On the set D we consider the following differential inclusion:

$$
(\dot{x},\dot{y}) \in \begin{cases} \{(-x-y,x+y)\} & \text{for } (x,y) \in \text{ int } D_1 \cup \text{ int } D_3;\\ \{(x-y,x-y)\} & \text{for } (x,y) \in \text{ int } D_2 \cup \text{ int } D_4;\\ \{(-x-y,x+y),(x-y,x-y)\} & \text{for } (x,y) \in D \cap \{(x,y):\ xy=0\}. \end{cases}
$$

We denote the right-hand side of the considered differential inclusion by G. Clearly, G is a uniformly bounded compact-valued upper semicontinuous mapping. Let us fix an arbitrary point (x_0,y_0) from the set D. Then there exists a trajectory of the considered differential inclusion starting from (x_0,y_0) and defined on the interval $[0,+\infty)$. Moreover, this trajectory is a periodic function with respect to time with period 4.

To show how Theorem 5.1 can be applied in this case, we set $F(x,y,z) := G(x,y)\times\{1\}$ and consider the differential inclusion $(\dot{x},\dot{y},\dot{z}) \in F(x,y,z)$ on the set $\hat{D} = D\times[0,+\infty)$.

Let k be an arbitrary positive integer. We are going to define a relatively open partitioning $\mathcal{U}^k = \{U_s^k,\ s=0,1,2,\ldots\}$ of $\hat{D}$. Let us fix arbitrarily the positive integer k and the nonnegative integer s. Let us denote by m and r the quotient and the remainder of s over $8k$, respectively, i.e., $s = 8mk+r$, where $r \in \{0,1,\ldots,8k-1\}$. We set U_s^k to be the set

$$
\left\{(x,y)\in D_1:\ \frac{r}{k}\leq x+y<\frac{r+1}{k},\ m\leq z<m+\frac{y}{x+y}\right\}
$$

$$
\text{if } r=0,1,\ldots,k-1;
$$

$$
\left\{(x,y)\in D_2:\ \frac{r}{k}-1\leq y-x<\frac{r+1}{k}-1,\ m\leq z<m+\frac{x}{x-y}\right\}
$$

$$
\text{if } r=k,k+1,\ldots,2k-1;
$$

$$
\left\{(x,y)\in D_3:\ \frac{r}{k}-2\leq -x-y<\frac{r+1}{k}-2,\ m\leq z<m+\frac{y}{x+y}\right\}
$$

$$
\text{if } r=2k,2k+1,\ldots,3k-1;
$$

$$\left\{(x,y)\in D_4:\ \frac{r}{k}-3\le x-y<\frac{r+1}{k}-3,\ m\le z<m+\frac{x}{x-y}\right\}$$
$$\text{if } r=3k,3k+1,\dots,4k-1;$$

$$\left\{(x,y)\in D_1:\ \frac{r}{k}-4\le x+y<\frac{r+1}{k}-4,\ m+\frac{y}{x+y}\le z<m+1\right\}$$
$$\text{if } r=4k,4k+1,\dots,5k-1;$$

$$\left\{(x,y)\in D_2:\ \frac{r}{k}-5\le y-x<\frac{r+1}{k}-5,\ m+\frac{x}{x-y}\le z<m+1\right\}$$
$$\text{if } r=5k,5k+1,\dots,6k-1;$$

$$\left\{(x,y)\in D_3:\ \frac{r}{k}-6\le -x-y<\frac{r+1}{k}-6,\ m+\frac{y}{x+y}\le z<m+1\right\}$$
$$\text{if } r=6k,6k+1,\dots,7k-1;$$

$$\left\{(x,y)\in D_4:\ \frac{r}{k}-7\le x-y<\frac{r+1}{k}-7,\ m+\frac{x}{x-y}\le z<m+1\right\}$$
$$\text{if } r=7k,7k+1,\dots,8k-1.$$

On the set $\hat{D}$ we define F_k as follows:

$$F_k(x,y,z)=\begin{cases}\{(-x-y,x+y,1)\} & \text{for } (x,y)\in D_1\cup D_3;\\ \{(x-y,x-y,1)\} & \text{for } (x,y)\in D_2\cup D_4.\end{cases}$$

It could be directly checked that the diameter of $F_k(U_s^k)$ is less than or equal to $\frac{\sqrt{2}}{k}$, and $F_{k+1}=F_k$ and $F_k\subset F$ for each k and each s. One can directly verify that for each point x_0 of $\hat{D}$, for each $T>0$, and for each positive integer k there exists an absolutely continuous function $x_k:[0,T]\to\hat{D}$ such that

$$\left|\begin{array}{l}\dot{x}(t)\in F_k(x(t)) \ \text{ a.e. on } [0,T],\\ x(0)=x_0.\end{array}\right. \tag{5.3}$$

Moreover, for each positive integer k and m with $k\ge m$, and for each solution $x_k(\cdot)$ of (5.1), the function

$$t\to s(t),\ \text{where } x_k(t)\in U^m_{s(t)},$$

is monotone increasing. Hence, Theorem 5.1 can be applied.

REFERENCES

[1] F. Ancona and A. Bressan, *Patchy vector fields and asymptotic stabilization*, ESAIM Control Optim. Calc. Var., 4 (1999), pp. 445–471.

[2] J.-P. Aubin and A. Cellina, *Differential Inclusions. Set-Valued Maps and Viability Theory*, Grundlehren Math. Wiss. 264, Springer-Verlag, Berlin, 1984.

[3] H. Bounkhel and T. Haddad, *Existence of viable solutions for nonconvex differential inclusions*, Electron. J. Differential Equations, 2005 (2005), pp. 1–10.
[4] A. Bressan, *Upper and lower semicontinuous case. A unified approach*, in Nonlinear Controllability and Optimal Control, H. Sussmann, ed., Marcel Dekker, New York, 1980, pp. 23–31.
[5] A. Bressan, *On differential relations with lower continuous right-hand side: An existence theorem*, J. Differential Equations, 37 (1980), pp. 89–97.
[6] A. Bressan, *Solutions of lower semicontinuous differential inclusions on closed sets*, Rend. Sem. Mat. Univ. Padova, 69 (1983), pp. 99–107.
[7] A. Bressan, *On the qualitative theory of lower semicontinuous differential inclusions*, J. Differential Equations, 77 (1989), pp. 379–391.
[8] A. Bressan and A. Cortesi, *Directionally continuous selections in Banach spaces*, Nonlinear Anal., 13 (1989), pp. 987–992.
[9] A. Bressan, A. Cellina, and G. Colombo, *Upper semicontinuous differential inclusions without convexity*, Proc. Amer. Math. Soc., 106 (1989), pp. 771–775.
[10] A. Cellina and A. Ornelas, *Existence of solutions to differential inclusions and to time optimal control problems in the autonomous case*, SIAM J. Control Optim., 42 (2003), pp. 260–265.
[11] F. Clarke, Y. Ledyaev, R. Stern, and P. Wolenski, *Nonsmooth Analysis and Control Theory*, Springer-Verlag, New York, 1998.
[12] A. Daniilidis and P. Georgiev, *Cyclic hypomonotonicity, cyclic submonotonicity and integration*, J. Optim. Theory Appl., 122 (2004), pp. 19–39.
[13] K. Deimling, *Multivalued Equations*, Walter de Gruyter, Berlin, New York, 1992.
[14] A. F. Filippov, *Differential equations with discontinuous right-hand side*, Amer. Math. Soc. Transl. Ser. 2, 42 (1964), pp. 199–231.
[15] A. F. Filippov, *The existence of solutions of generalized differential equations*, Math. Notes, 10 (1971), pp. 608–611.
[16] A. F. Filippov, *Differential Equations with Discontinuous Right-Hand Sides*, Math. Appl. (Soviet Ser.) 18, F. M. Arscott, ed., Kluwer Academic, Dordrecht, The Netherlands, 1988 (transl. from the Russian).
[17] M. K. Fort, *Points of continuity of semi-continuous functions*, Publ. Math. Debrecen, 2 (1951), pp. 100–102.
[18] H. Hermes, *The generalized differential equation $\dot{x} \in R(t, x)$*, Adv. Math., 4 (1970), pp. 149–169.
[19] H. Kaczynski and C. Olech, *Existence of solutions of orientor fields with non-convex right-hand side*, Ann. Polon. Math., 29 (1974), pp. 61–66.
[20] S. Lojasiewicz, *The existence of solutions for lower semicontinuous orientor fields*, Bull. Acad. Polon. Sci. Sér. Sci. Math., 28 (1980), pp. 483–487.
[21] S. Lojasiewicz, *Some theorems of Scorza Dragoni type for multifunctions with application to the problem of existence of solutions for differential multivalued equations*, in Mathematical Control Theory, Banach Center Publ. 14, PWN, Warsaw, 1985, pp. 625–643.
[22] C. Olech, *Existence of solutions of non-convex orientor fields*, Boll. Un. Mat. Ital. (4), 11 (1975), pp. 189–197.
[23] N. Ribarska, *Internal characterization of fragmentable spaces*, Mathematika, 34 (1987), pp. 243–257.
[24] V. V. Srivatsa, *Baire class* 1 *selectors for upper semicontinuous set-valued maps*, Trans. Amer. Math. Soc., 337 (1993), pp. 609–624.
[25] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Grundlehren Math. Wiss. 317, Springer-Verlag, Berlin, 1998.
[26] V. Veliov, *Stability-like properties of differential inclusions*, Set-Valued Anal., 5 (1997), pp. 73–88.

SIAM J. OPTIM.
Vol. 18, No. 3, pp. 752–777

VARIATIONAL ANALYSIS OF EVOLUTION INCLUSIONS*

BORIS MORDUKHOVICH†

Abstract. The paper is devoted to optimization problems of the Bolza and Mayer types for evolution systems governed by nonconvex Lipschitzian differential inclusions in Banach spaces under endpoint constraints described by finitely many equalities and inequalities with generally nonsmooth functions. We develop a variational analysis of such problems mainly based on their discrete approximations and the usage of advanced tools of generalized differentiation satisfying comprehensive calculus rules in the framework of Asplund (and hence any reflexive Banach) spaces. In this way we establish extended results on stability of discrete approximations (with the strong $W^{1,2}$-convergence of optimal solutions under consistent perturbations of endpoint constraints) and derive necessary optimality conditions for nonconvex discrete-time and continuous-time systems in the refined Euler–Lagrange and Weierstrass–Pontryagin forms accompanied by the appropriate transversality inclusions. In contrast to the case of geometric endpoint constraints in infinite dimensions, the necessary optimality conditions obtained in this paper do not impose any nonempty interiority/finite-codimension/normal compactness assumptions. The approach and results developed in the paper make a bridge between optimal control/dynamic optimization and constrained mathematical programming problems in infinite-dimensional spaces.

Key words. variational analysis, dynamic optimization and optimal control, evolution and differential inclusions, Banach and Asplund spaces, discrete/finite-difference approximations, nondifferentiable programming, generalized differentiation, necessary optimality conditions

AMS subject classifications. 49J53, 49J52, 49J24, 49M25, 90C30

DOI. 10.1137/060652889

1. Introduction.

This paper concerns the study of dynamic optimization problems governed by constrained evolution systems in infinite-dimensional spaces. We pay attention mainly to variational analysis of the following *generalized Bolza problem* (P) for differential inclusions in Banach spaces with endpoint constraints described by finitely many equalities and inequalities.

Let X be a Banach *state space* with the *initial state* $x_0 \in X$, and let $T := [a, b] \subset \mathbb{R}$ be a fixed *time interval*. Given a set-valued mapping $F\colon X \times T \rightrightarrows X$ and real-valued functions $\varphi_i\colon X \to \mathbb{R}$ as $i = 0, \ldots, m+r$ and $\vartheta\colon X \times X \times T \to \mathbb{R}$, consider the following problem:

$$\text{minimize} \quad J[x] := \varphi_0\big(x(b)\big) + \int_a^b \vartheta\big(x(t), \dot{x}(t), t\big)\, dt \tag{1.1}$$

subject to *dynamic constraints* governed by the evolution/differential inclusion

$$\dot{x}(t) \in F\big(x(t), t\big) \quad \text{a.e.} \quad t \in [a, b] \quad \text{with} \quad x(a) = x_0 \tag{1.2}$$

with *functional endpoint constraints* of the inequality and equality types given by

$$\varphi_i\big(x(b)\big) \le 0, \quad i = 1, \ldots, m, \tag{1.3}$$

*Received by the editors February 24, 2006; accepted for publication (in revised form) November 14, 2006; published electronically October 4, 2007. This research was partially supported by the U.S. National Science Foundation under grants DMS-0304989 and DMS-0603846 and by the Australian Research Council under grant DP-0451168.

http://www.siam.org/journals/siopt/18-3/65288.html

†Department of Mathematics, Wayne State University, Detroit, MI 48202 (boris@math.wayne.edu).

$$\varphi_i\big(x(b)\big) = 0, \quad i = m+1, \ldots, m+r. \tag{1.4}$$

Note that $\dot{x}(t)$ stands in (1.1) for the time derivative of $x(t)$ and that "a.e." (almost everywhere) signifies as usual that the inclusion holds up to the Lebesgue measure zero on $\mathbb{R}$. The initial state x_0 and the time interval T are fixed in problem (P) for simplicity; the methods developed in this paper allow us to include $x(a)$ and $[a, b]$ in the optimization process and to derive necessary optimality conditions for these variable data.

Dynamic optimization problems for differential inclusions with the *finite-dimensional* state space $X = \mathbb{R}^n$ have been intensively studied over the years, especially during the last decade, mainly from the viewpoint of deriving necessary optimality conditions; see [3, 8, 12, 14, 17, 19] for various results, methods, and more references. Dynamic optimization problems governed by infinite-dimensional *evolution equations* have also been much investigated, motivated mainly by applications to optimal control of partial differential equations; see, e.g., the books [7, 10] and the references therein. To the best of our knowledge, deriving necessary optimality conditions in dynamic optimization problems for evolution systems governed by differential inclusions in *infinite-dimensional spaces* has not drawn attention in the literature till very recently.

In the book [14], the author developed the method of *discrete approximations* to study optimal control problems of minimizing the Bolza functional (1.1) over appropriate solutions to evolution systems governed by infinite-dimensional differential inclusions of type (1.2) with endpoint constrains given in the *geometric form*

$$x(b) \in \Omega \subset X \tag{1.5}$$

via closed subsets of Banach spaces satisfying certain requirements. The major assumption on Ω made in [14] is the so-called *sequential normal compactness* (SNC) property of Ω at the optimal endpoint $\bar{x}(b) \in \Omega$; see [13] for a comprehensive theory for this and related properties, which play a significant role in infinite-dimensional variational analysis and its applications. Loosely speaking, the SNC property means that Ω should be "sufficiently fat" around the reference point; e.g., it never holds for singletons unless X is finite-dimensional, where the SNC property is satisfied for every nonempty set. For *convex* sets in infinite-dimensional spaces, the SNC property automatically holds when $\operatorname{int}\Omega \neq \emptyset$. Furthermore, it happens to be closely related [14] to the so-called finite-codimension property of convex sets, which is known to be essential for the fulfillment of an appropriate counterpart of the Pontryagin maximum principle for infinite-dimensional systems of optimal control; see the books by Fattorini [7] and by Li and Yong [10] for the corresponding results, discussions, counterexamples, and more references.

In this paper we show that the dynamic optimization problem (P) formulated above, with the *functional* endpoint constraints (1.3) and (1.4) given by *finitely many* Lipschitz continuous functions on a broad class of Banach spaces (that particularly includes every reflexive space), admits necessary optimality conditions in the extended Euler–Lagrange form accompanied by the corresponding Weierstrass–Pontryagin/maximum and transversality relations with *no SNC* and similar assumptions imposed on the underlying endpoint constraint set. Moreover, the case of endpoint constraints (1.3) and (1.4) under consideration allows us to partly avoid some other rather restrictive assumptions (like "strong coderivative normality," which may not hold in infinite-dimensional spaces; see sections 6 and 7 for more details) imposed in [14] in the general case of geometric constraints (1.5). Our approach is based, in

addition to the results of [14], on certain delicate properties of appropriate *subdifferentials* of locally Lipschitzian functions on infinite-dimensional spaces, as well as on *dual/coderivative characterizations* of Lipschitzian and metric regularity properties of set-valued mappings.

The rest of the paper is organized as follows. In section 2 we formulate the standing assumptions on the initial data of (P), make more precise the solution concept for the evolution inclusion (1.1) and the types of local minimizers to (P) under consideration, and also discuss the relaxation procedure used for some results and proofs in the paper. Our main focus in this paper is on the so-called *intermediate local minimizers*, which occupy (strictly) an intermediate position between the classical weak and strong minima, being nevertheless closer to strong minimizers from the viewpoint of necessary optimality conditions for differential inclusions.

In section 3 we construct a sequence of the *well-posed discrete approximations* (P_N) to the original Bolza problem (P), which take into account specific features of the functional endpoint constraints (1.3) and (1.4) involving *consistent perturbations* of these constraints in the discrete approximation procedure. Then we present a major result on the *strong stability* of discrete approximations that justifies the $W^{1,2}$-norm convergence of optimal solutions for (P_N) to the fixed local minimizer for the original problem (P).

Section 4 contains an overview of the basic tools of *generalized differentiation* needed to perform the subsequent variational analysis of the discrete-time and continuous-time evolution systems under consideration in infinite-dimensional spaces. Most of the material in this section is taken from the author's book [13], where the reader can find more results and commentaries in this direction and related topics.

Section 5 is devoted to deriving necessary optimality conditions for the constrained *discrete-time* problems arising from the discrete approximation procedure whose well-posedness and stability are justified in section 3. These problems are reduced to (nondynamic) constrained problems of mathematical programming in infinite dimensions, which happen to be *intrinsically nonsmooth* and involve finitely many functional and geometric constraints generated by those in (1.2)–(1.4) via the discrete approximation procedure. Variational analysis of such problems requires applications of the full power of the generalized differential calculus in infinite-dimensional spaces developed in [13].

In section 6 we derive necessary optimality conditions of the extended *Euler–Lagrange* type for *relaxed* intermediate minimizers to the original Bolza problem (P) by passing to the limit from those obtained for discrete-time problems in section 5. It is worth emphasizing that the realization of the limiting procedure requires not only the strong convergence of optimal trajectories to discrete approximation problems established in section 3 but also justifying an appropriate convergence of *adjoint trajectories* in necessary optimality conditions for the sequence of discrete-time inclusions. The latter becomes passable due to specific properties of the basic generalized differential constructions reviewed in section 4, which include complete *dual characterizations* of Lipschitzian and metric regularity properties of set-valued mappings.

The concluding section 7 concerns necessary optimality conditions for arbitrary (*nonrelaxed*) intermediate minimizers to problem (P), considering for simplicity the Mayer form (P_M) with no integral term in (1.1), that are established in terms of the *extended Euler–Lagrange* inclusion accompanied by the *Weierstrass–Pontryagin/maximum* and transversality relations without imposing any SNC assumptions on the target/endpoint constraint set described by (1.3) and (1.4). The approach is

based on an additional approximation procedure that allows us to reduce (P_M) to an unconstrained (while nonsmooth and nonconvex) Bolza problem of the type treated in section 6, for which *any* intermediate local minimizer happens to be a relaxed one. The passage to the limit from the latter approximation is largely similar to that developed in section 6, not requiring, however, any relaxation requirement due to the usage of *Ekeland's variational principle*.

Our notation is basically standard; cf. [13, 14]. Unless otherwise stated, all the spaces considered are Banach with the norm $\|\cdot\|$ and the canonical dual pairing $\langle\cdot,\cdot\rangle$ between the space in question, say X, and its topological dual X^* whose weak* topology is denoted by w^*. We use the symbols $\mathbb{B}$ and $\mathbb{B}^*$ to signify the closed unit balls of the space under consideration and its dual, respectively. Given a set-valued mapping $F\colon X \rightrightarrows X^*$, its *sequential Painlevé–Kuratowski upper/outer limit* at $\bar{x}$ is defined by

$$\operatorname*{Lim\,sup}_{x\to\bar{x}} F(x) := \Big\{x^* \in X^* \;\Big|\; \exists \text{ sequences } x_k \to \bar{x},\ x_k^* \overset{w^*}{\to} x^* \text{ with } x_k^* \in F(x_k) \text{ as } k \in \mathbb{N} := \{1,2,\ldots\}\Big\}. \tag{1.6}$$

2. The generalized Bolza problem for evolution inclusions. For the sake of brevity and simplicity, we consider in this paper the Bolza problem (P) with *autonomous* (time-independent) data, i.e., when $\vartheta = \vartheta(x,v)$ in (1.1) and $F = F(x)$ in (1.2). The case of nonautonomous systems can be studied in a manner similar to that of [14, Chapter 6], which is devoted to problems with geometric constraints of type (1.5). Let us start with the precise definition of solutions (trajectories, arcs) to the differential inclusion (1.2) following the book by Deimling [6].

Definition 2.1 (solutions to differential inclusions in infinite-dimensional spaces). *By a* solution *to inclusion* (1.2) *we understand a mapping $x\colon T \to X$, which is Fréchet differentiable for a.e. $t \in T$ satisfying* (1.2) *and the* Newton–Leibniz formula

$$x(t) = x_0 + \int_a^t \dot{x}(s)\,ds \quad \textit{for all } \ t \in T,$$

where the integral in taken in the Bochner sense.

It is well known that for $X = \mathbb{R}^n$, $x(t)$ is a.e. differentiable on T and satisfies the Newton–Leibniz formula *if and only if* it is *absolutely continuous* on T in the standard sense. However, for infinite-dimensional spaces X even the Lipschitz continuity may not imply the a.e. differentiability. On the other hand, there is a *complete characterization* of Banach spaces X, where the absolute continuity of every $x\colon T \to X$ is *equivalent* to its a.e. differentiability and the fulfillment of the Newton–Leibniz formula: this is the class of spaces having the so-called *Radon–Nikodým property* (RNP), which is well investigated in the geometric theory of Banach spaces [4]. Observe, in particular, that every *reflexive* space enjoys the RNP.

Recall further that a Banach space X is *Asplund* if any of its separable subspaces has a separable dual. This is a major subclass of Banach spaces that particularly includes every space with a *Fréchet differentiable renorm* off the origin (i.e., every *reflexive* space), every space with a separable dual, etc.; see [4] for more details, characterizations, and references. There is a deep relationship between spaces having the RNP and Asplund spaces, which is used in what follows: *given a Banach space X, the dual space X^* has the RNP if and only if X is Asplund.*

It has been well recognized that differential inclusions (1.2), which are certainly interesting on their own, provide a useful generalization of *control systems* governed

by differential/evolution *equations* with control parameters:

$$\dot{x} = f(x,u), \quad u \in U, \tag{2.1}$$

where the control sets $U(\cdot)$ may also depend on time and *state* variables via $F(x,t) = f(x,U(x,t),t)$. In some cases, especially when the sets $F(\cdot)$ are convex, the differential inclusions (1.2) admit parametric representations of type (2.1), but in general they cannot be reduced to parametric control systems and should be studied for their own sake. Note also that the *ODE form* (2.1) in Banach spaces is strongly related to various control problems for evolution *partial differential equations* of parabolic and hyperbolic types, where solutions may be understood in some other appropriate senses; see, e.g., the books [7, 10, 14] for more discussions.

In what follows, we focus our attention on the study of *intermediate local minimizers* for problem (P) introduced by the author in [12]. Recall that a *feasible arc* to (P) is a solution to the differential inclusion (1.2), in the sense of Definition 2.1, for which $J[x] < \infty$ in (1.1) and the endpoint constraints (1.3) and (1.4) are satisfied.

DEFINITION 2.2 (intermediate local minimizers). *A feasible arc $\bar{x}(\cdot)$ is an* intermediate local minimizer *of rank $p \in [1,\infty)$ for (P) if there are numbers $\epsilon > 0$ and $\alpha \ge 0$ such that $J[\bar{x}] \le J[x]$ for any feasible arcs to (P) satisfying the relationships*

$$\|x(t) - \bar{x}(t)\| < \epsilon \ \text{ for all } \ t \in [a,b] \quad \text{and} \tag{2.2}$$

$$\alpha \int_a^b \|\dot{x}(t) - \dot{\bar{x}}(t)\|^p \, dt < \epsilon. \tag{2.3}$$

In fact, relationships (2.2) and (2.3) mean that we consider a neighborhood of $\bar{x}(\cdot)$ in the Sobolev space $W^{1,p}\big([a,b];X\big)$ with the norm

$$\|x(\cdot)\|_{W^{1,p}} := \max_{t\in[a,b]} \|x(t)\| + \Big(\int_a^b \|\dot{x}(t)\|^p \, dt \Big)^{1/p},$$

where the norm on the right-hand side is taken in the space X. If there is only the requirement (2.2) in Definition 2.2, i.e., $\alpha = 0$ in (2.3), then we get the classical *strong* local minimum corresponding to a neighborhood of $\bar{x}(\cdot)$ in the norm topology of $C\big([a,b];X\big)$. If instead of (2.3) one puts the more restrictive requirement

$$\|\dot{x}(t) - \dot{\bar{x}}(t)\| < \epsilon \ \text{ a.e. } \ t \in [a,b],$$

then we have the classical *weak* local minimum in the framework of Definition 2.2. Thus the introduced notion of intermediate local minimizers takes, for any $p \in [1,\infty)$, an *intermediate* position between the classical concepts of strong ($\alpha = 0$) and weak ($p = \infty$) local minima, being indeed different from both classical notions; see various examples in [20, 14]. Clearly all the necessary conditions for intermediate local minimizers automatically hold for strong (and hence for global) minimizers.

Considering the autonomous Bolza problem (P) in this paper, we impose the following *standing assumptions* on its initial data along a given intermediate local minimizer $\bar{x}(\cdot)$:

(H1) There are an open set $U \subset X$ and a number $\ell_F > 0$ such that $\bar{x}(t) \in U$ for all $t \in [a,b]$, and the sets $F(x)$ are nonempty and compact for all $x \in U$ and satisfy the inclusion

$$F(x) \subset F(u) + \ell_F \|x-u\| \mathbb{B} \ \text{ whenever } \ x,u \in U, \tag{2.4}$$

which implies the uniform boundedness of the sets $F(x)$ on U, i.e., the existence of some constant $\gamma > 0$ such that

$$F(x) \subset \gamma\mathbb{B} \quad \text{for all} \quad x \in U.$$

(H2) The integrand ϑ is Lipschitz continuous on $U \times (\gamma\mathbb{B})$.

(H3) The endpoint functions φ_i, $i = 0, \ldots, m+r$, are locally Lipschitzian around $\bar{x}(b)$ with the common Lipschitz constant $\ell > 0$.

Observe that (2.4) is equivalent to saying that the set-valued mapping F is *locally Lipschitzian* around $\bar{x}(\cdot)$ with respect to the classical Hausdorff metric on the space of nonempty and compact subsets of X.

In what follows, along with the original problem (P), we consider its "relaxed" counterpart significantly used in some results and proofs of the paper. Roughly speaking, the relaxed problem is obtained from (P) by a *convexification* procedure with respect to the *velocity* variable. It follows the route of Bogolyubov and Young in the classical calculus of variations and of Gamkrelidze and Warga in optimal control; see the book [14] and the references therein for more details and commentaries.

To construct an appropriate relaxation of the Bolza problem (P) under consideration, we first consider the extended-real-valued function

$$\vartheta_F(x, v) := \vartheta(x, v) + \delta\big(v; F(x)\big),$$

where $\delta(\cdot; \Omega)$ is the *indicator function* of the set Ω equal to 0 on Ω and to ∞ out of it. Denote by

$$\widehat{\vartheta}_F(x, v) := \big(\vartheta_F\big)_v^{**}(x, v), \quad (x, v) \in X \times X,$$

the *biconjugate/bipolar* function to $\vartheta_F(x, \cdot)$, i.e., the greatest proper, convex, and lower semicontinuous function with respect to v, which is majorized by ϑ_F. Then the *relaxed problem* (R) to (P), or the *relaxation* of (P), is defined as follows:

$$\text{minimize} \quad \widehat{J}[x] := \varphi_0\big(x(b)\big) + \int_a^b \widehat{\vartheta}_F\big(x(t), \dot{x}(t)\big)\, dt \tag{2.5}$$

over a.e. differentiable arcs $x\colon [a, b] \to X$ that are Bochner integrable on $[a, b]$ together with $\vartheta_F\big(x(t), \dot{x}(t)\big)$, satisfying the Newton–Leibniz formula and the endpoint constraints (1.3), (1.4).

Note that the feasibility requirement $\widehat{J}[x] < \infty$ in (2.5) is fulfilled only if $x(\cdot)$ is a solution (in the sense of Definition 2.1) to the *convexified differential inclusion*

$$\dot{x}(t) \in \operatorname{clco} F\big(x(t), \dot{x}(t)\big) \ \text{ a.e. } \ t \in [a, b] \ \text{ with } \ x(a) = x_0, \tag{2.6}$$

where "clco" stands for the convex closure of a set in X. Thus the relaxed problem (R) can be considered under explicit dynamic constraints given by the convexified differential inclusion (2.6). Any trajectory for (2.6) is called a *relaxed trajectory* for (1.2), in contrast to the *ordinary* (or *original*) trajectories for the latter inclusion.

Deep relationships exist between relaxed and ordinary trajectories for differential inclusions, which reflect the fundamental *hidden convexity* inherent in continuous-time (nonatomic measure) dynamic systems defined by differential and integral operators. In particular, *any relaxed trajectory* of (1.2) under assumption (H1) can be *uniformly approximated* (in the $C\big([a, b]; X\big)$-norm) by a sequence of ordinary trajectories; see,

e.g., [6, 18]. We need the following version [5] of this approximation/density property involving not only differential inclusions but also minimizing functionals.

LEMMA 2.3 (approximation property for the relaxed Bolza problem). *Let $x(\cdot)$ be a relaxed trajectory for the differential inclusion* (1.2) *with a separable state space X, where F and ϑ satisfy assumptions* (H1) *and* (H2)*, respectively. Then there is sequence of the ordinary trajectories $x_k(\cdot)$ for* (1.2) *such that $x_k(\cdot) \to x(\cdot)$ in $C\big([a,b];X\big)$ as $k \to \infty$ and*

$$\liminf_{k\to\infty} \int_a^b \vartheta\big(x_k(t),\dot{x}_k(t)\big)\,dt \le \int_a^b \widehat{\vartheta}_F\big(x(t),\dot{x}(t)\big)\,dt.$$

Note that Lemma 2.3 does *not* assert that the approximating trajectories $x_k(\cdot)$ satisfy the endpoint constraints (1.3) and (1.4). Indeed, there are examples showing that the latter may not be possible and, moreover, the property of *relaxation stability*

$$\inf(P) = \inf(R) \tag{2.7}$$

is violated; in (2.7) the infima of the cost functionals (1.1) and (2.5) are taken over all the feasible arcs in (P) and (R), respectively.

An obvious sufficient condition for the relaxation stability is the *convexity* of the sets $F(x,t)$ and of the integrand ϑ in v. However, the relaxation stability goes *far beyond* the standard convexity due to the *hidden convexity* property of continuous-time differential systems. In particular, Lemma 2.3 ensures the relaxation stability of nonconvex problems (P) with no constraints on the endpoint $x(b)$. There are various efficient conditions for the relaxation stability of nonconvex problems with endpoint and other constraints; see [14, subsection 6.1.2] with the commentaries therein for more details, discussions, and references.

A *local* version of the relaxation stability property (2.7) regarding intermediate minimizers for the Bolza problem (P) is postulated as follows.

DEFINITION 2.4 (relaxed intermediate local minimizers). *A feasible arc $\bar{x}(\cdot)$ to the Bolza problem (P) is a* relaxed intermediate local minimizer *of rank $p \in [1,\infty)$ for (P) if it is an intermediate local minimizer of this rank for the relaxed problem (R) providing the same value of the cost functionals: $J[\bar{x}] = \widehat{J}[\bar{x}]$.*

It is not hard to observe that, under the standing assumptions formulated above, the notions of intermediate local minima and relaxed intermediate local minima do not actually depend on rank p, i.e., they either hold or violate for all $p \in [1,\infty)$ simultaneously. In what follows we always take (unless otherwise stated in section 7) $p = 2$ and $\alpha = 1$ in (2.3) for simplicity.

The principal method of our study in this paper involves *discrete approximations* of the original Bolza problem (P) for constrained continuous-time evolution inclusions by a family of dynamic optimization problems of Bolza type governed by discrete-time inclusions with endpoint constraints. We show that this method generally leads to necessary optimality conditions for *relaxed* intermediate local minimizers of (P). Then an additional approximation procedure allows us to establish necessary optimality conditions for *arbitrary* (nonrelaxed) intermediate local minimizers by reducing them to problems which are *automatically* stable with respect to relaxation.

3. Stability of discrete approximations. In this section we present basic constructions of the method of discrete approximations in the theory of necessary optimality conditions for differential inclusions following the scheme of [12, 14] developed for the case of geometric constraints, with certain modifications required for the func-

tional endpoint constraints (1.3) and (1.4) under consideration in infinite-dimensional spaces.

Since we use discrete approximations mostly from a "theoretical" viewpoint (as a vehicle to derive necessary optimality conditions), in what follows we use just the simplest finite-difference replacement of the derivative by the *uniform Euler scheme*:

$$\dot{x}(t) \approx \frac{x(t+h)-x(t)}{h}, \quad h \to 0.$$

To formalize this process, we take any natural number $N \in \mathbb{N}$ and consider the *discrete mesh* on T defined by

$$T_N := \{a, a+h_N, \ldots, b-h_N, b\}, \quad h_N := (b-a)/N,$$

with the *stepsize of discretization* h_N and the *mesh points* $t_j := a + jh_N$ as $j = 0, \ldots, N$, where $t_0 = a$ and $t_N = b$. Then the differential inclusion (1.2) is replaced by a sequence of its *discrete approximations*

$$x_N(t_{j+1}) \in x_N(t_j) + h_N F\big(x_N(t_j)\big), \;\; j = 0, \ldots, N-1, \quad x(t_0) = x_0. \tag{3.1}$$

Given a discrete trajectory $x_N(t_j)$ satisfying (3.1), we consider its *piecewise linear extension* $x_N(t)$ to the continuous-time interval $T = [a, b]$, i.e., the *Euler broken lines*. We also define the *piecewise constant extension* to T of the corresponding *discrete velocity* by

$$v_N(t) := \frac{x_N(t_{j+1}) - x_N(t_j)}{h_N}, \quad t \in [t_j, t_{j+1}), \;\; j = 0, \ldots, N-1.$$

It follows from the very definition of the Bochner integral that

$$x_N(t) = x_0 + \int_a^t v_N(s)\, ds \;\; \text{for} \;\; t \in T.$$

The next result, which plays a significant role in the method of discrete approximations, establishes the *strong $W^{1,2}$-norm approximation* of *any* trajectory for the differential inclusion (1.2) by extended trajectories of the sequence of discrete inclusions (3.1) under the general assumptions made in (H1). Note that the norm convergence in $W^{1,2}\big([a, b]; X\big)$ implies the *uniform* convergence of the trajectories on $[a, b]$ and the *pointwise*, for a.e. $t \in [a, b]$, convergence of (some subsequence of) their *derivatives*. The latter is crucial for the purposes of this paper, especially in the case of *nonconvex-valued* differential inclusions. The proof of this result in given in [14, Theorem 6.4], which is an infinite-dimensional counterpart of the one in [12, Theorem 3.1].

LEMMA 3.1 (strong $W^{1,2}$-approximation by discrete trajectories). *Let $\bar{x}(\cdot)$ be an arbitrary solution to the differential inclusion* (1.2) *under the assumptions in* (H1)*, where X is a general Banach space. Then there is a sequence of solutions $\widehat{x}_N(t_j)$ to the discrete inclusions* (3.1) *such that their extensions $\widehat{x}_N(t)$, $a \le t \le b$, converge to $\bar{x}(t)$ strongly in the space $W^{1,2}\big([a, b]; X\big)$ as $N \to \infty$.*

Observe that the proof of the above result given in [12, 14] is *constructive* and provides efficient estimates of the convergence rate being of certain independent interest for numerical analysis.

Now fix an *intermediate local minimizer* $\bar{x}(\cdot)$ for the Bolza problem (P) and construct a sequence of discrete approximation problems (P_N), $N \in \mathbb{N}$, admitting

optimal solutions $\bar{x}_N(\cdot)$ whose extensions converge to $\bar{x}(\cdot)$ in the norm topology of $W^{1,2}\big([a,b];X\big)$ as $N\to\infty$.

To proceed, we take a sequence of the discrete trajectories $\widehat{x}_N(\cdot)$ approximating by Lemma 3.1 the given local minimizer $\bar{x}(\cdot)$ to (P) and denote

$$\eta_N := \max_{j\in\{1,\ldots,N\}} \|\widehat{x}_N(t_j)-\bar{x}(t_j)\| \to 0 \ \text{ as } \ N\to\infty. \tag{3.2}$$

In [14, subsection 6.1.1], the reader can find more information on computing and estimating η_N, which is not needed in what follows: it is sufficient to know that $\eta_N\to 0$ as $N\to\infty$.

Having $\epsilon>0$ from relations (2.2) and (2.3) for the intermediate minimizer $\bar{x}(\cdot)$ with $p=2$ and $\alpha=1$, we always suppose that

$$\bar{x}(t)+\epsilon/2\in U \ \text{ for all } \ t\in[a,b],$$

where U is a neighborhood of $\bar{x}(\cdot)$ from (H1). Let $\ell>0$ be the common Lipschitz constant of φ_i, $i=1,\ldots,m+r$, from (H3). Construct problems (P_N), $N\in\mathbb{N}$, as follows: minimize

$$\begin{aligned} J_N[x_N] := {}& \varphi_0\big(x_N(t_N)\big)+h_N\sum_{j=0}^{N-1}\vartheta\Big(x_N(t_j),\frac{x_N(t_{j+1})-x_N(t_j)}{h_N}\Big)\\ &+\sum_{j=0}^{N-1}\int_{t_j}^{t_{j+1}}\Big\|\frac{x_N(t_{j+1})-x_N(t_j)}{h_N}-\dot{\bar{x}}(t)\Big\|^2\,dt \end{aligned} \tag{3.3}$$

over discrete trajectories $x_N=x_N(\cdot)=\big(x_0,x_N(t_1),\ldots,x_N(t_N)\big)$ for the difference inclusions (3.1) subject to the constraints

$$\varphi_i\big(x_N(t_N)\big)\le\ell\eta_N \ \text{ for } \ i=1,\ldots,m, \tag{3.4}$$

$$-\ell\eta_N\le\varphi_i\big(x_N(t_N)\big)\le\ell\eta_N \ \text{ for } \ i=m+1,\ldots,m+r, \tag{3.5}$$

$$\|x_N(t_j)-\bar{x}(t_j)\|\le\frac{\epsilon}{2} \ \text{ for } \ j=1,\ldots,N, \quad \text{and} \tag{3.6}$$

$$\sum_{j=0}^{N-1}\int_{t_j}^{t_{j+1}}\Big\|\frac{x_N(t_{j+1})-x_N(t_j)}{h_N}-\dot{\bar{x}}(t)\Big\|^2\,dt\le\frac{\epsilon}{2}. \tag{3.7}$$

Considering in what follows (without further mentioning) the piecewise linear extension of $x_N(\cdot)$ to the whole interval $[a,b]$, we observe the relationships:

$$\begin{cases} x_N(t)=x_0+\displaystyle\int_a^t\dot{x}_N(s)\,ds \quad \text{for all } t\in[a,b] \quad \text{and}\\[2ex] \dot{x}_N(t)=\dot{x}_N(t_j)\in F\big(x_N(t_j)\big), \ \ t\in[t_j,t_{j+1}), \ \ j=0,\ldots,N-1. \end{cases} \tag{3.8}$$

In the next theorem, we establish that the given *relaxed* intermediate local minimizer $\bar{x}(\cdot)$ to (P) can be approximated by *optimal solutions* to (P_N) *strongly* in $W^{1,2}\big([a,b];X\big)$; the latter implies the a.e. *pointwise* convergence of the derivatives

significant for the main results of the paper. To justify such an approximation, we need to impose the *Asplund* structure on *both* the state space X and its dual X^*, which is particularly the case when X is *reflexive*. Note also there are *nonreflexive* (even separable) spaces for which both X and X^* are Asplund; see, e.g., [4].

THEOREM 3.2 (strong convergence of discrete optimal solutions). *Let $\bar{x}(\cdot)$ be a relaxed intermediate local minimizer for the Bolza problem (P) under the standing assumptions* (H1)–(H3) *in the Banach state space X, and let (P_N), $N \in \mathbb{N}$, be a sequence of discrete approximation problems built above. The following hold:*

(i) *Each (P_N) admits an optimal solution.*

(ii) *If in addition both X and X^* are Asplund, then any sequence $\{\bar{x}_N(\cdot)\}$ of optimal solutions to (P_N) converges to $\bar{x}(\cdot)$ strongly in $W^{1,2}\big([a,b];X\big)$.*

Proof. To justify assertion (i), we first observe that the set of *feasible* solutions to each problem (P_N) is *nonempty* for all $N \in \mathbb{N}$ sufficiently large. Indeed, pick the discrete trajectory $\widehat{x}_N(\cdot)$ approximating the given local minimizer $\bar{x}(\cdot)$ by Lemma 3.1 and show that it satisfies all the constraints (3.4)–(3.7) for large N. By assumption (H3) we have

$$\big|\varphi_i\big(\widehat{x}_N(t_N)\big) - \varphi_i\big(\bar{x}(b)\big)\big| \le \ell \,\|\widehat{x}(t_N) - \bar{x}(t_N)\| \le \ell\eta_N \;\text{ for all }\; i = 1,\dots,m+r$$

due to (3.2). This implies the fulfillment of the endpoint constraints (3.4) and (3.5) for $\widehat{x}_N(\cdot)$, since those in (1.3) and (1.4) hold for $\bar{x}(\cdot)$. The fulfillment of (3.6) for $\widehat{x}_N(\cdot)$ follows directly from the construction of $\eta_N \to 0$ in (3.2). Further, it is easy to check that

$$\sum_{j=0}^{N-1} \int_{t_j}^{t_{j+1}} \Big\| \frac{\widehat{x}_N(t_{j+1}) - \widehat{x}_N(t_j)}{h_N} - \dot{\bar{x}}(t) \Big\|^2 dt = \int_a^b \|\dot{\widehat{x}}_N(t) - \dot{\bar{x}}(t)\|^2 \, dt =: \alpha_N$$

for the piecewise linear extension of $\widehat{x}_N(\cdot)$ to $[a,b]$. By the $W^{1,2}$-approximation in Lemma 3.1 we have that $\alpha_N \to 0$ as $N \to \infty$, which justifies the fulfillment of (3.7) for large N. The *existence of optimal solutions* to (P_N) follows now from the classical Weierstrass theorem due to the compactness and continuity assumptions made in (H1)–(H3).

Let us now prove the convergence assertion (ii) under the additional assumptions on the state space. Check first the *value convergence*

$$J_N[\widehat{x}_N] \to J[\bar{x}] \quad \text{as} \quad N \to \infty \tag{3.9}$$

along a subsequence of $N \to \infty$. Considering the expression for $J_N[\widehat{x}_N]$ in (3.3) and using assumptions (H2) and (H3), we observe that (3.9) follows from

$$\begin{aligned} h_N \sum_{j=0}^{N-1} \vartheta\Big(\widehat{x}_N(t_j), \frac{\widehat{x}_N(t_{j+1}) - \widehat{x}_N(t_j)}{h_N}\Big) &= \sum_{j=0}^{N-1} \int_{t_j}^{t_{j+1}} \vartheta\big(\widehat{x}_N(t_j), \dot{\widehat{x}}_N(t)\big)\, dt \\ = \int_a^b \vartheta\big(\widehat{x}_N(t), \dot{\widehat{x}}_N(t)\big)\, dt + O(h_N) &\to \int_a^b \vartheta\big(\bar{x}(t), \dot{\bar{x}}(t)\big)\, dt \;\text{ as }\; N \to \infty, \end{aligned}$$

which hold by Lemma 3.1 ensuring the a.e. convergence $\dot{\widehat{x}}_N(t) \to \dot{\bar{x}}(t)$ along a subsequence and by the Lebesgue dominated convergence theorem valid for the Bochner integral.

None of the previous arguments used either the relaxation property of the intermediate minimizer or the Asplund property of X and X^*. Now we are going to

employ these properties to justify the relationship

$$\lim_{N\to\infty}\Big[\beta_N:=\int_a^b\|\dot{\bar{x}}_N(t)-\dot{\bar{x}}(t)\|^2\,dt\Big]=0 \tag{3.10}$$

for *every* sequence of optimal solutions $\bar{x}_N(\cdot)$ to (P_N).

Arguing by contradiction, pick a limiting point $\beta>0$ of $\{\beta_N\}$ in (3.10) and suppose for simplicity that $\beta_N\to\beta$ for all $N\to\infty$. To proceed, observe that both spaces X and X^* *enjoy the RNP*. Indeed, the one for X^* is equivalent to the Asplund property of X, while the Asplund property of X^* ensures the RNP for X due to the latter fact and that of $X\subset X^{**}$. Taking into account (H1) and (3.8), we apply to the sequence $\{\dot{\bar{x}}_N(\cdot)\}$ the Dunford theorem [4, Theorem IV.1] on the *weak compactness* in $L^1\big([a,b];X\big)$, which allows us to find a subsequence of $\{\dot{\bar{x}}_N(\cdot)\}$ and a mapping $v(\cdot)\in L^1\big([a,b];X\big)$ such that

$$\dot{\bar{x}}_N(\cdot)\to v(\cdot)\ \text{ weakly in }\ L^1\big([a,b];X\big)\ \text{ as }\ N\to\infty. \tag{3.11}$$

Using (3.8) and the compactness in $C\big([a,b];X\big)$ of solution sets for differential inclusions that holds under the assumptions made in (H1) (see, e.g., [18, Theorem 3.4.2]), we conclude that the sequence $\{\bar{x}_N(\cdot)\}$ contains a subsequence that converges to some $\widetilde{x}\in C\big([a,b];X\big)$ in the norm topology of the space $C\big([a,b];X\big)$. Passing to the limit in the first relationship of (3.8), while taking into account (3.11) and the weak continuity of the Bochner integral as an operator from $L^1\big([a,b];X\big)$ to X, we arrive at the representation

$$\widetilde{x}(t)=x_0+\int_a^t v(s)\,ds\ \text{ for all }\ t\in[a,b],$$

which implies that $v(t)=\dot{\widetilde{x}}(t)$ for a.e. $t\in[a,b]$.

Furthermore, the classical Mazur *weak closure* theorem ensures that $\widetilde{x}(\cdot)$ is a solution to the *convexified* differential inclusion (2.6). By the structure of problems (P_N) and by the construction of $\widetilde{x}(\cdot)$, it is not hard to conclude that $\widetilde{x}(\cdot)$ satisfies the endpoint constraints (1.3) and (1.4) and that it belongs to the prescribed ϵ-neighborhood of $\bar{x}(\cdot)$ in the norm topology of $W^{1,2}\big([a,b];X\big)$. By passing to the limit in the obvious inequality

$$J_N[\bar{x}_N]\le J_N[\widehat{x}_N]\ \text{ for all large }\ N\in\mathbb{N},$$

while taking into account (3.9) and the lower semicontinuity of the *convexified* integrand $\widehat{\vartheta}_F(x,\cdot)$ from (2.5) in the *weak* topology of $L^2\big([a,b];X\big)$, we get

$$\widehat{J}[\widetilde{x}]=\varphi_0\big(\widetilde{x}(b)\big)+\int_a^b\widehat{\vartheta}_F\big(\widetilde{x}(t),\dot{\widetilde{x}}(t)\big)\,dt+\beta\le J[\bar{x}].$$

Since $\beta>0$ and $J[\bar{x}]=\widehat{J}[\bar{x}]$, the latter gives $\widehat{J}[\widetilde{x}]<\widehat{J}[\bar{x}]$, which contradicts the choice of $\bar{x}(\cdot)$ as a relaxed intermediate local minimizer for (P). Thus (3.10) holds, and so $\bar{x}_N(\cdot)\to\bar{x}(\cdot)$ as $N\to\infty$ strongly in $W^{1,2}\big([a,b];X\big)$. This completes the proof of the theorem. □

The strong convergence result of Theorem 3.2 *makes a bridge* between the original continuous-time dynamic optimization problem (P) and its discrete-time counterparts (P_N), which allows us to derive necessary optimality conditions for (P) by passing to the limit from those for (P_N). The latter ones are *intrinsically nonsmooth* and require appropriate tools of generalized differentiation for their variational analysis.

4. Generalized differentiation. In this section, we define the main constructions of generalized differentiation used in what follows. Since the major framework in this paper is the class *Asplund spaces*, we present simplified definitions and some properties held in this setting. All the material reviewed and employed below is taken from the author's book [13], where the reader can find more details and references.

We start with generalized normals to closed sets $\Omega \subset X$. Given $\bar{x} \in \Omega$, the (basic, limiting) *normal cone* to Ω at $\bar{x}$ is defined by

$$N(\bar{x};\Omega) := \operatorname*{Lim\,sup}_{x\to\bar{x}} \widehat{N}(x;\Omega), \tag{4.1}$$

where "Lim sup" stands for the *sequential* upper/outer limit (1.6) of the *Fréchet normal cone* (or the *prenormal cone*) to Ω at $x \in \Omega$ given by

$$\widehat{N}(x;\Omega) := \left\{ x^* \in X^* \Big| \limsup_{u \overset{\Omega}{\to} x} \frac{\langle x^*, u-x\rangle}{\|u-x\|} \le 0 \right\}, \tag{4.2}$$

where $x \overset{\Omega}{\to} \bar{x}$ signifies that $x \to \bar{x}$ with $x \in \Omega$, and where $\widehat{N}(x;\Omega) := \emptyset$ for $x \notin \Omega$.

Given a set-valued mapping $F\colon X \rightrightarrows Y$ of closed graph

$$\operatorname{gph} F := \big\{(x,y) \in X \times Y \big| \; y \in F(x)\big\},$$

define its *normal coderivative* and *Fréchet coderivative* at $(\bar{x},\bar{y}) \in \operatorname{gph} F$ by, respectively,

$$D^*F(\bar{x},\bar{y})(y^*) := \big\{x^* \in X^* \big| \; (x^*,-y^*) \in N\big((\bar{x},\bar{y});\operatorname{gph} F\big)\big\}, \tag{4.3}$$

$$\widehat{D}^*F(\bar{x},\bar{y})(y^*) := \big\{x^* \in X^* \big| \; (x^*,-y^*) \in \widehat{N}\big((\bar{x},\bar{y});\operatorname{gph} F\big)\big\}. \tag{4.4}$$

If $F = f\colon X \to Y$ is *strictly differentiable* at $\bar{x}$ (in particular, if $f \in C^1$), then

$$D^*f(\bar{x})(y^*) = \widehat{D}^*f(\bar{x})(y^*) = \big\{\nabla f(\bar{x})^* y^*\big\}, \quad y^* \in Y^*,$$

i.e., both coderivatives (4.3) and (4.4) are positively homogeneous extensions of the classical *adjoint* derivative operator to nonsmooth and set-valued mappings.

Finally, consider a function $\varphi\colon X \to \mathbb{R}$ *locally Lipschitzian* around $\bar{x}$; in this paper we do not use more general functions. Then the (basic, limiting) *subdifferential* of φ at $\bar{x}$ is defined by

$$\partial\varphi(\bar{x}) := \operatorname*{Lim\,sup}_{x\to\bar{x}} \widehat{\partial}\varphi(x), \tag{4.5}$$

where the sequential outer limit (1.6) of the *Fréchet subdifferential* mapping $\widehat{\partial}\varphi(\cdot)$ is given by

$$\widehat{\partial}\varphi(x) := \left\{ x^* \in X^* \Big| \; \frac{\varphi(u)-\varphi(x)-\langle x^*, u-x\rangle}{\|u-x\|} \ge 0 \right\}. \tag{4.6}$$

We are not going to review in this section appropriate properties of the generalized differential constructions (4.1)–(4.6) used in sections 5–7; these properties will be invoked with the exact references to [13] in the corresponding places of the proofs in the subsequent sections. Just note here that our basic/limiting constructions (4.1),

(4.3), and (4.5) enjoy *full calculus* in the framework of Asplund spaces, while the Fréchet-like ones (4.2), (4.4), and (4.6) satisfy certain rules of "fuzzy calculus." Both of these calculi are employed in what follows. The reader can find some additional and related material in the books by Rockafellar and Wets [16], Smirnov [17], and Vinter [19] (concerning exact/full calculus in finite dimensions) and in the book by Borwein and Zhu [1] on fuzzy calculus in infinite dimensions; see also the references therein.

5. Optimality conditions for discrete inclusions. In this section we derive necessary optimality conditions for the sequence of discrete approximation problems (P_N) defined in (3.1) and (3.3)–(3.7). We present results only in the "fuzzy" form, which are more convenient for deriving necessary conditions for the original problem (P) by the limiting procedure in section 6. "Pointwise" necessary conditions for (P_N) and for related discrete-time problems (not used in this paper) can be found in [14, subsection 6.1.4].

Observe first that each discrete optimization problem (P_N) can be equivalently written in a special form of *constrained mathematical programming* problem in infinite-dimensional spaces:

$$\text{(5.1)}\qquad \begin{cases} \text{minimize } \ \psi_0(z) \ \text{ subject to} \\ \psi_j(z) \le 0, \quad j = 1, \dots, s, \\ f(z) = 0, \\ z \in \Theta_j \subset Z, \quad j = 1, \dots, l, \end{cases}$$

where ψ_j are real-valued functions on some Banach space Z, where $f\colon Z \to E$ is a mapping between Banach spaces, and where $\Theta_j \subset Z$. To see this, we let

$$z := (x_1, \dots, x_N, v_0, \dots, v_{N-1}) \in Z := X^{2N},$$
$$E := X^N, \ s := N + 2 + m + 2r, \ l := N - 1$$

and rewrite (P_N) as a mathematical programming problem (5.1) with the following data:

$$\text{(5.2)}\qquad \psi_0(z) := \varphi_0(x_N) + h_N \sum_{j=0}^{N-1} \vartheta(x_j, v_j) + \sum_{j=0}^{N-1} \int_{t_j}^{t_{j+1}} \|v_j - \dot{\bar{x}}(t)\|^2 \, dt,$$

$$\text{(5.3)}\qquad \psi_j(z) := \begin{cases} \|x_{j-1} - \bar{x}(t_{j-1})\| - \epsilon/2, \quad j = 1, \dots, N+1, \\ \displaystyle\sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \|v_i - \dot{\bar{x}}(t)\|^2 \, dt - \epsilon/2, \quad j = N+2, \\ \varphi_i(x_N) - \ell\eta_N, \quad j = N+2+i, \quad i = 1, \dots, m+r, \\ -\varphi_i(x_N) - \ell\eta_N, \quad j = N+2+m+r+i, \quad i = 1, \dots, r, \end{cases}$$

$$\text{(5.4)}\qquad \begin{cases} f(z) = \big(f_0(z), \dots, f_{N-1}(z)\big) \ \text{ with} \\ f_j(z) := x_{j+1} - x_j - h_N v_j, \quad j = 0, \dots, N-1, \end{cases}$$

$$\Theta_j := \Big\{ z \in X^{2N} \Big| \; v_j \in F(x_j) \Big\} \;\; \text{for} \;\; j = 0, \ldots, N-1 \tag{5.5}$$

in terms of the initial data of problem (P_N).

The next theorem establishes necessary optimality conditions for each problem (P_N) in the *approximate/fuzzy* form of refined *Euler–Lagrange* and *transversality inclusions* expressed in terms of Fréchet-like normals and subgradients. The proof is based on applying the corresponding *fuzzy calculus* rules and *neighborhood criteria* for *metric regularity and Lipschitzian behavior* of mappings from [13]. Note that fuzzy calculus rules provide representations of Fréchet subgradients and normals of sums and intersections at the reference points via those at points that are arbitrarily close to the reference ones. *Just for notational simplicity*, we suppose in the formulation and proof of the next theorem that these *arbitrarily close points reduce to the reference points in question*. This convention does not restrict the generality from the viewpoint of our main goal to derive necessary optimality conditions in the continuous-time problem (P). Indeed, the possible difference between the mentioned points obviously disappears in the limiting procedure.

THEOREM 5.1 (fuzzy Euler–Lagrange conditions for discrete approximations). *Let $\bar{x}_N(\cdot) = \{\bar{x}_N(t_j) | \; j = 0, \ldots, N\}$ be local optimal solutions to problems (P_N) as $N \to \infty$ under the assumptions* (H1)–(H3) *with the Asplund state space X. Consider the quantities*

$$\theta_{Nj} := 2 \int_{t_j}^{t_{j+1}} \Big\| \frac{\bar{x}_N(t_{j+1}) - \bar{x}_N(t_j)}{h_N} - \dot{\bar{x}}(t) \Big\| \, dt, \quad j = 0, \ldots, N-1. \tag{5.6}$$

Then there exists a sequence $\varepsilon_N \downarrow 0$ along some $N \to \infty$, and there are sequences of Lagrange multipliers λ_{iN}, $i = 0, \ldots, m+r$, and adjoint trajectories $p_N(\cdot) = \{p_N(t_j) \in X^* | \; j = 0, \ldots, N\}$ satisfying the following relationships:

- *the sign and nontriviality conditions*

$$\lambda_{iN} \ge 0 \;\; \textit{for all} \;\; i = 0, \ldots, m+r, \quad \sum_{i=0}^{m+r} \lambda_{iN} = 1; \tag{5.7}$$

- *the complementary slackness conditions*

$$\lambda_{iN} \Big[\varphi_i\big(\bar{x}_N(t_N)\big) - \ell\eta_N \Big] = 0 \;\; \textit{for} \;\; i = 1, \ldots, m; \tag{5.8}$$

- *the extended Euler–Lagrange inclusion in the approximate form*

$$\begin{aligned} \Big(\frac{p_N(t_{j+1}) - p_N(t_j)}{h_N}, p_N(t_{j+1}) - \lambda_{0N} \frac{\theta_{Nj}}{h_N} b^*_{N_j} \Big) &\in \lambda_{0N} \widehat{\partial} \vartheta \Big(\bar{x}_N(t_j), \frac{\bar{x}_N(t_{j+1}) - \bar{x}_N(t_j)}{h_N} \Big) \\ &+ \widehat{N} \Big(\Big(\bar{x}_N(t_j), \frac{\bar{x}_N(t_{j+1}) - \bar{x}_N(t_j)}{h_N} \Big); \operatorname{gph} F \Big) + \varepsilon \mathbb{B}^* \;\; \textit{with} \;\; b^*_{Nj} \in \mathbb{B}^*, \quad j = 0, \ldots, N-1; \end{aligned} \tag{5.9}$$

- *the approximate transversality inclusion*

$$\begin{aligned} -p_N(t_N) \in \sum_{i=0}^{m} \lambda_{iN} \widehat{\partial} \varphi_i\big(\bar{x}_N(t_N)\big) + \sum_{i=m+1}^{m+r} \lambda_{iN} \Big[\widehat{\partial} \varphi_i\big(\bar{x}_N(t_N)\big) \bigcup \widehat{\partial}\big(-\varphi_i\big)\big(\bar{x}_N(t_N)\big) \Big] \\ + \varepsilon \mathbb{B}^*. \end{aligned} \tag{5.10}$$

Proof. Consider problem (P_N) for any fixed $N \in N$ in the equivalent mathematical programming form (5.1) with the initial data (5.2)–(5.5). Denote

$$\bar{z} := \big(\bar{x}_N(t_1), \ldots, \bar{x}_N(t_N), \bar{v}_N(t_0), \ldots, \bar{v}_N(t_{N-1})\big)$$

and take N so large that the $W^{1,2}$-constraints (3.6) and (3.7) for $\bar{x}_N(\cdot)$ hold with the strict inequality, which is possible by Theorem 3.2. Thus the latter constraints can be simply *ignored* in what follows.

To prove the theorem, it is sufficient to examine the following two mutually exclusive cases, which completely cover the situation.

Case 1. Assume that the operator constraint mapping $f\colon X^{2N} \to X^N$ in (5.1) and (5.4) is *metrically regular* at $\bar{z}$ relative to the set $\Theta := \Theta_0 \cap \cdots \cap \Theta_{N-1}$ in (5.5) in the sense that there is a constant $\mu > 0$ and a neighborhood U of $\bar{z}$ satisfying the distance estimate

$$\operatorname{dist}(z; S) \le \mu \|f(z) - f(\bar{z})\| \ \text{ for all } \ z \in \Theta \cap U, \ \text{ where } \ S := \big\{z \in \Theta \big|\ f(z) = f(\bar{z})\big\}.$$

Then applying Ioffe's *exact penalization theorem* (see [14, Theorem 5.16]) and taking into account the specific structures of the inequality constraint functions ψ_j in (5.3) for $j = N + 2 + i$ as $i = 1, \ldots, m + r$, we conclude that $\bar{z}$ is a local optimal solution to the *unconstrained penalized problem*:

$$\begin{aligned} \text{minimize} \quad & \max\big\{\psi_0(z) - \psi_0(\bar{z}),\ \max_{i \in I(\bar{x}_N)} \varphi_i(x_N)\big\} \\ & + \ \mu\big(\|f(z)\| + \operatorname{dist}(z; \Theta)\big) \end{aligned} \tag{5.11}$$

for all $\mu > 0$ sufficiently large, where

$$I(\bar{x}_N) := \big\{i \in \{1, \ldots, m\}\big|\ \varphi_i(\bar{x}_N) = \ell\eta_N\big\}$$

$$\bigcup \big\{i \in \{m+1, \ldots, m+r\}\big|\ \text{either } \varphi_i(\bar{x}_N) = \ell\eta_N \text{ or } -\varphi_i(\bar{x}_N) = \ell\eta_N\big\}.$$

Applying the generalized Fermat rule [13, Proposition 1.114] to the local optimal solution $\bar{z}$ for (5.11), we arrive at the subdifferential inclusion

$$0 \in \widehat{\partial}\Big[\max\big\{\psi_0(\cdot) - \psi_0(\bar{z}),\ \max_{i \in I(\bar{x}_N)} \varphi_i(\cdot)\big\} + \mu\|f(\cdot)\| + \mu \operatorname{dist}(\cdot; \Theta)\Big](\bar{z}). \tag{5.12}$$

Fix any $\varepsilon > 0$ and apply the fuzzy sum rule for Fréchet subgradients from [13, Theorem 2.33(b)] to (5.12). This gives (remember out notational convention)

$$0 \in \widehat{\partial}\Big[\max\big\{\psi_0(\cdot) - \psi_0(\bar{z}),\ \max_{i \in I(\bar{x}_N)} \varphi_i(\cdot)\big\}\Big](\bar{z}) + \mu\,\widehat{\partial}\|f(\cdot)\|(\bar{z}) + \mu\,\widehat{\partial}\operatorname{dist}(\bar{z}; \Theta) + (\varepsilon/4)\mathbb{B}^*.$$

Computing now by [13, Proposition 1.85] the Fréchet subdifferential of the distance function $\operatorname{dist}(\bar{z}; \Theta)$ and using the simple chain rule for the composition $\|f(z)\| = (\phi \circ f)(z)$ with $\phi(y) := \|y\|$ and the smooth mapping f from (5.4), we get

$$0 \in \widehat{\partial}\Big[\max\big\{\psi_0(\cdot) - \psi_0(\bar{z}),\ \max_{i \in I(\bar{x}_N)} \varphi_i(\cdot)\big\}\Big](\bar{z}) + \sum_{j=0}^{N-1} \nabla f_j(\bar{z})^* e_j^* + \widehat{N}(\bar{z}; \Theta) + (\varepsilon/4)\mathbb{B}^*$$

for some $e_j^* \in X^*$ with, by the structure of f in (5.4),

$$\begin{aligned} \sum_{j=0}^{N-1} \nabla f_j(\bar{z})^* e_j^* = \big(& -e_0^*, e_0^* - e_1^*, \ldots, e_{N-2}^* \\ & - e_{N-1}^*, e_{N-1}^*, -h_N e_0^*, \ldots, -h_N e_{N-1}^*\big). \end{aligned} \tag{5.13}$$

To proceed further, we use the fuzzy intersection rule from [13, Lemma 3.1] ensuring that

$$\widehat{N}(\bar{z};\Theta)\subset\widehat{N}(\bar{z};\Theta_0)+\cdots+\widehat{N}(\bar{z};\Theta_{N-1})+(\varepsilon/4)\mathbb{B}^*$$

and also employ the fuzzy rule for Fréchet subgradients of the maximum function (cf. [13, Theorem 3.46] and its proof), giving, by the structure of the index set $I(\bar{x}_N)$, the inclusion

$$\begin{aligned}&\widehat{\partial}\Big[\max\big\{\psi_0(\cdot)-\psi_0(\bar{z}),\max_{i\in I(\bar{x}_N)}\varphi_i(\cdot)\big\}\Big](\bar{z})\\&\subset\sum_{i=0}^{m}\lambda_{iN}\widehat{\partial}\varphi_i(\bar{x}_N)+\sum_{i=m+1}^{m+r}\lambda_{iN}\Big[\widehat{\partial}\varphi_i(\bar{x}_N)\bigcup\widehat{\partial}(-\varphi_i)(\bar{x}_N)\Big]+(\varepsilon/4)\mathbb{B}^*,\end{aligned}$$

where the multipliers λ_{iN}, $i=0,\ldots,m+r$, satisfy the sign, nontriviality, and complementary slackness conditions in (5.7) and (5.8).

Applying the aforementioned fuzzy sum rule to the cost function (5.2) and taking into account the classical relationship

$$\partial\|\cdot\|^2(x)\subset 2\|x\|\mathbb{B}^*\ \text{ for any }\ x\in X,$$

as well as the subdifferentiation formula under the integral sign in (5.2) well known from convex analysis, we have the inclusion

$$\widehat{\partial}\psi_0(\bar{z})\subset\widehat{\partial}\varphi_0(\bar{x}_N)+h_N\sum_{j=0}^{N-1}\Big[\widehat{\partial}\vartheta\big(\bar{x}_j,\bar{v}_j\big)+\big(0,2\theta_{Nj}\mathbb{B}^*\big)\Big]+(\varepsilon/4)\mathbb{B}^*,$$

where θ_{Nj} are defined in (5.6). Finally, choose $p_N(t_0)\in X^*$ arbitrarily and let

$$p_N(t_j):=e^*_{j-1},\quad j=1,\ldots,N,$$

with e^*_j given in (5.13). Then taking into account the special separated structures of the sets Θ_j in (5.5), we arrive at the Euler–Lagrange and transversality inclusions (5.9) and (5.10) with $\varepsilon_N=\varepsilon$ by substituting the corresponding fuzzy relationships derived above into the generalized Fermat stationary condition (5.12). This completes the proof of the theorem in Case 1.

Case 2. It remains to consider the situation when the mapping f from (5.4) is *not metrically regular* at $\bar{z}$ relative to the set intersection $\Theta:=\Theta_0\cap\cdots\cap\Theta_{N-1}$. Let us show that this *never holds*, along some subsequences $\varepsilon_N\downarrow 0$ and $N\to\infty$, under the local Lipschitzian assumption imposed on F.

Indeed, in this case the *restriction* $f_\Theta\colon X^{2N}\to X^N$ of f to Θ defined by

$$f_\Theta(z):=\begin{cases}f(z) & \text{if }\ z\in\Theta,\\ \emptyset & \text{otherwise}\end{cases}$$

is obviously *not* metrically regular *around* $\bar{z}$ in the sense of [13, Definition 1.47]. Then the *characterization* of the latter property from [13, Theorem 4.5] allows us, for any fixed $\varepsilon>0$, to find $z\in\bar{z}+\varepsilon\mathbb{B}$ and $e^*=(e^*_0,\ldots,e^*_{N-1})\in(X^*)^N$ such that

$$\|e^*\|>1\ \text{ and }\ 0\in\widehat{D}^*f_\Theta(z)(e^*)$$

via the Fréchet coderivative (4.4) of f_Θ. Employing now the coderivative sum rule from [13, Theorem 1.62] and the fuzzy intersection rule from [13, Lemma 3.1], we get

$$0 \in \sum_{j=0}^{N-1} \nabla f_j(z)^* e_j^* + \sum_{j=0}^{N-1} \widehat{N}(z_j;\Theta_j) + \varepsilon\mathbb{B}^*$$

with some $z_j \in \Theta_j \cap (z+\varepsilon\mathbb{B})$. According to our notation agreement, we put $z_j = z = \bar{z}$ for simplicity. Thus there are $z_j^* \in \widehat{N}(\bar{z};\Theta_j)$ satisfying

$$-\sum_{j=0}^{N-1} z_j \in \sum_{j=0}^{N-1} \nabla f_j(\bar{z})^* e_j^* + \varepsilon\mathbb{B}^*. \tag{5.14}$$

Taking into account calculation (5.13) due to the form (5.4) of the mapping f and the specific structures of the sets Θ_j in (5.5), we find from (5.14) dual elements

$$(x_{ij}^*, v_{ij}^*) \in \widehat{N}\left(\Big(\bar{x}_N(t_j), \frac{\bar{x}_N(t_{j+1}) - \bar{x}_N(t_j)}{h_N}\Big); \operatorname{gph} F_j\right), \quad j = 0,\ldots,N-1,$$

satisfying the relationships

$$\begin{cases} -x_{jj}^* - e_{j-1}^* + e_j^* \in \varepsilon\mathbb{B}^*, \quad j = 0,\ldots,N-1, \\ -v_{jj}^* + h_N e_j^* \in \varepsilon\mathbb{B}^*, \quad j = 0,\ldots,N-1, \\ -e_{N-1}^* \in \varepsilon\mathbb{B}^*. \end{cases}$$

Define now the adjoint discrete trajectory $p_N(t_j)$, $j = 0,\ldots,N$, as in Case 1. Then the above relationships ensure that the pair $\big(\bar{x}_N(\cdot), p_N(\cdot)\big)$ satisfies the Euler–Lagrange inclusion (5.9) and the transversality inclusion (5.10) with

$$\lambda_{iN} = 0 \ \text{ for all } \ i = 0,\ldots,m+r$$

and the following nontriviality condition:

$$\|p_N(t_1)\| + \cdots + \|p_N(t_N)\| \ge 1 \ \text{ for all large } \ N \in \mathbb{N}. \tag{5.15}$$

Let us show that condition (5.15) *contradicts* (5.9) and (5.10) with $\lambda_{iN} = 0$ due the *locally Lipschitzian* property of F assumed in the theorem.

To proceed, we observe that the Euler–Lagrange inclusion (5.9) with $\lambda_{0N} = 0$ can be equivalently written as

$$\frac{p_N(t_{j+1}) - p_N(t_j)}{h_N} \in \widehat{D}^* F\Big(\bar{x}_N(t_j), \frac{\bar{x}_N(t_{j+1}) - \bar{x}_N(t_j)}{h_N}\Big)\big(-p_N(t_{j+1})\big) + \varepsilon\mathbb{B}^*, \quad j = 0,\ldots,N-1.$$

Then the local Lipschitzian property of F with modulus ℓ_F yields, by the *neighborhood coderivative characterization* of [13, Theorem 4.7], that

$$\|x_j^*\| \le \ell_F \|v_j^*\| \ \text{ whenever } \ x_j^* \in \widehat{D}^* F_j(x_j, v_j)(v_j^*)$$

and (x_j, v_j) near $\big(\bar{x}_N(t_j), [\bar{x}_N(t_{j+1}) - \bar{x}_N(t_j)]/h_N\big)$. Thus

$$\|p_N(t_{N-1})\| \le \|p_N(t_N)\|\big(1 + h_N \ell_F\big) + \varepsilon h_N$$

and then, as a discrete counterpart of the Gronwall lemma,

$$\|p_N(t_j)\| \le \exp\big(\ell_F(b-a)\big)\|p_N(t_N)\| + \varepsilon(b-a) \ \text{ for all } \ j=0,\dots,N. \tag{5.16}$$

Finally, take a sequence $\nu_k \downarrow 0$ as $k\to\infty$ and choose numbers N_k and ε_k such that

$$N_k := \big[1/\nu_k\big] \ \text{ and } \ \varepsilon_k \le \nu_k^2 \ \text{ as } \ k\in\mathbb{N},$$

where $[\cdot]$ stands for the greatest integer less than or equal to the given real number. By (5.16) and by the transversality condition (5.10) with $\lambda_{iN}=0$ along the chosen sequences of $\varepsilon_k=\varepsilon_{N_k}\downarrow 0$ and $N_k\to\infty$ as $k\to\infty$, we have the estimate

$$\sum_{j=1}^{N_k}\|p_{N_k}(t_j)\| \le \nu_k\exp\big(\ell_F(b-a)\big)+\nu_k(b-a)\downarrow 0 \ \text{ as } \ k\in\mathbb{N},$$

which contradicts (5.15) and completes the proof of the theorem. □

6. Extended Euler–Lagrange conditions for relaxed minimizers. In this section we derive necessary optimality conditions in the refined forms of the extended Euler–Lagrange and transversality inclusions for *relaxed* intermediate local minimizers of the original problem (P). The proof is based on the passing to the limit from the necessary optimality conditions for discrete approximation problems obtained in section 5 and on the usage of the *strong stability* of discrete approximations established in section 3. A crucial part of the proof involves the justification of an appropriate convergence of *adjoint arcs*; the latter becomes possible due to the *coderivative characterization* of Lipschitzian set-valued mappings taken from [13].

THEOREM 6.1 (extended Euler–Lagrange and transversality inclusions for relaxed intermediate minimizers). *Let $\bar{x}(\cdot)$ be a relaxed intermediate local minimizer for the Bolza problem (P) given in (1.1)–(1.4) under the standing assumptions of section 2, where the spaces X and X^* are Asplund. Then there are nontrivial Lagrange multipliers $0\ne(\lambda_0,\dots,\lambda_{m+r})\in\mathbb{R}^{m+r+1}$ and an absolutely continuous mapping $p\colon[a,b]\to X^*$ such that the following necessary conditions hold:*

- *the sign conditions*

$$\lambda_i\ge 0 \ \text{ for all } \ i=0,\dots,m+r; \tag{6.1}$$

- *the complementary slackness conditions*

$$\lambda_i\varphi_i\big(\bar{x}(b)\big)=0 \ \text{ for } \ i=1,\dots,m; \tag{6.2}$$

- *the extended Euler–Lagrange inclusion, for a.e., $t\in[a,b]$,*

$$\dot{p}(t)\in \operatorname{clco}\Big\{u\in X^*\Big|\ \big(u,p(t)\big)\in\lambda_0\partial\vartheta\big(\bar{x}(t),\dot{\bar{x}}(t)\big)+N\big((\bar{x}(t),\dot{\bar{x}}(t));\operatorname{gph}F\big)\Big\}; \tag{6.3}$$

- *and the transversality inclusion*

$$-p(b)\in\sum_{i=0}^{m}\lambda_i\partial\varphi_i\big(\bar{x}(b)\big)+\sum_{i=m+1}^{m+r}\lambda_i\Big[\partial\varphi_i\big(\bar{x}(b)\big)\bigcup\partial\big(-\varphi_i\big)\big(\bar{x}(b)\big)\Big]. \tag{6.4}$$

Proof. Given the intermediate local minimizer $\bar{x}(\cdot)$ to (P), employ Theorem 3.2, which ensures the strong $W^{1,2}$-approximation of $\bar{x}(\cdot)$ by a sequence of optimal solutions $\bar{x}_N(\cdot)$ to problems (P_N). Applying now the necessary optimality condition

of Theorem 5.1, we find sequence of multipliers λ_{iN}, $i=0,\ldots,m+r$, and adjoint trajectories $p_N(\cdot)$ satisfying conditions (5.7)–(5.10). Without loss of generality, we can and do suppose that

$$\lambda_{iN}\to\lambda_i \ \text{ as } \ N\to\infty \ \text{ for all } \ i=0,\ldots,m+r,$$

where the limiting multipliers λ_i, $i=0,\ldots,m+r$, are not zero simultaneously and satisfy the sign condition (6.1). Moreover, we get the complementarity slackness conditions (6.2) by passing to the limit in (5.8) with $\eta_N\to 0$ as $N\to\infty$.

Let us next justify the possibility of passing to the limit in the approximate Euler–Lagrange (5.9) and transversality (5.10) inclusions for the discrete-time problems (P_N). Having θ_{Nj} defined in (5.6), consider the corresponding sequence of functions $\theta_N\colon[a,b]\to\mathbb{R}$ given by

$$\theta_N(t):=\frac{\theta_{Nj}}{h_N}b^*_{Nj} \ \text{ for } \ t\in[t_j,t_{j+1}), \ \ j=0,\ldots,N-1.$$

It follows from the strong $W^{1,2}$-convergence of Theorem 3.2 that

$$\begin{aligned}\int_a^b\big\|\theta_N(t)\big\|\,dt \ &\le \sum_{j=0}^{N-1}\theta_{Nj}\le 2\sum_{j=0}^{N-1}\int_{t_j}^{t_{j+1}}\Big\|\frac{\bar x_N(t_{j+1})-\bar x_N(t_j)}{h_N}-\dot{\bar x}_N(t)\Big\|\,dt\\ &=2\int_a^b\|\dot{\bar x}_N(t)-\dot{\bar x}(t)\|\,dt\to 0,\end{aligned}$$

which allows us to suppose without loss of generality that

$$\dot{\bar x}_N(t)\to\dot{\bar x}(t) \ \text{ and } \ \theta_N(t)\to 0 \ \text{ a.e. } \ t\in[a,b] \ \text{ as } \ N\to\infty.$$

Furthermore, we derive from the approximate Euler–Lagrange condition (5.9) that there are

$$e^*_{Nj},\widetilde e^*_{Nj}\in\mathbb{B}^* \ \text{ and } \ (x^*_{Nj},v^*_{Nj})\in\widehat\partial\vartheta\Big(\bar x_N(t_j),\frac{\bar x_N(t_{j+1})-\bar x_N(t_j)}{h_N}\Big), \quad j=0,\ldots,N-1,$$

such that the discrete-time inclusions

$$\begin{aligned}&\Big(\frac{p_N(t_{j+1})-p_N(t_j)}{h_N}-\lambda_{0N}x^*_{Nj}\Big)+\varepsilon_N e^*_{N_j}\\ &\in\widehat D^*F\Big(\bar x_N(t_j),\frac{\bar x_N(t_{j+1})-\bar x_N(t_j)}{h_N}\Big)\Big(\lambda_N v^*_{Nj}+\lambda_{0N}\frac{\theta_{Nj}}{h_N}b^*_{Nj}-p_N(t_{j+1})+\varepsilon_N\widetilde e^*_{N_j}\Big)\end{aligned}$$

hold for all $j=0,\ldots,N-1$ and all $N\in\mathbb{N}$. Observe that, due to [13, Proposition 1.85], the sequences $\{(x^*_{Nj},v^*_{jN})\}$ are uniformly bounded for all $j=0,\ldots,N-1$ by the Lipschitz constant of ϑ. Since the mapping F is locally Lipschitzian with constant ℓ_F, we get by the coderivative condition for the Lipschitz continuity from [13, Theorem 1.43] that

$$\Big\|\frac{p_N(t_{j+1})-p_N(t_j)}{h_N}-\lambda_N x^*_{Nj}+\varepsilon_N e^*_{N_j}\Big\|\le\ell_F\Big\|\lambda_N v^*_{Nj}+\lambda_N\frac{\theta_{Nj}}{h_N}b^*_{Nj}-p_N(t_{j+1})+\varepsilon_N\widetilde e^*_{N_j}\Big\|$$

for $j=0,\ldots,N-1$. This allows us to conclude that the piecewise extensions $p_N(t)$, $a\le t\le b$, of the adjoint discrete arcs $p_N(\cdot)$ are uniformly bounded on $[a,b]$ with

$$\|\dot p_N(t)\|\le\alpha+\beta\|\theta_N(t)\| \ \text{ a.e. } \ t\in[a,b], \tag{6.5}$$

where the positive numbers α and β are independent of N. Using now the Dunford criterion for the weak compactness in $L^1\big([a,b];X^*\big)$ from [4, Theorem IV.1] (note that both X and X^* enjoy the RNP due to their assumed Asplund property) and arguing similarly to the proof of Theorem 3.2 above, we find an absolute continuous mapping $p\colon[a,b]\to X^*$ satisfying the Newton–Leibniz formula and such that $\dot p_N(\cdot)\to\dot p(\cdot)$ as $N\to\infty$ (with no loss of generality) in the weak topology of $L^1\big([a,b];X^*\big)$.

Next we conclude from the approximate transversality inclusion (5.10), the sign and nontriviality conditions in (5.7), and the local Lipschitz continuity of φ_i, $i=0,\dots,m+r$, with the common constant ℓ from (H3) that

$$\|p_N(b)\|\le\ell(m+2)+1\ \text{ for sufficiently large }\ N\in\mathbb{N}$$

due to the uniform boundedness of Fréchet subgradients of locally Lipschitzian functions by [13, Proposition 1.85]. Since X is Asplund, this implies the weak* sequential compactness of $\{p_N(b)\}$ in X^*. Thus, passing to the limit in (5.10) as $N\to\infty$ and using definition (4.5) of the basic subdifferential, we arrive at the transversality inclusion (6.4).

Considering now the approximate Euler–Lagrange inclusion (5.9), we equivalently rewrite it as

$$\begin{aligned}\dot p_N(t)\in\Big\{u\in X^*\ \Big|\ \big(u,p_N(t_{j+1})-\lambda_{0N}\theta_N(t)\big)\in\lambda_{0N}\widehat\partial\vartheta(\bar x_N(t_j),\dot{\bar x}_N(t))\\ +\ \widehat N\big((\bar x_N(t_j),\dot{\bar x}_N(t)\big);\operatorname{gph}F\big)+\varepsilon_N\mathbb{B}^*\Big\}\end{aligned}\tag{6.6}$$

for $t\in[t_j,t_{j+1})$ with $j=0,\dots,N-1$. Observe, by the weak continuity of the Bochner integral in the Newton–Leibniz formula and by $\dot p_N(\cdot)\to\dot p(\cdot)$ weakly in $L^1\big[a,b];X^*\big)$, that the values $p_N(t)$ converge to $p(t)$ weakly in X^*. Furthermore, the Mazur weak closure theorem ensures that some sequence of *convex combinations* of $\{\dot p_N(\cdot)\}$ converges to $\dot p(\cdot)$ strongly in $L^1\big([a,b];X^*\big)$ as $N\to\infty$, and hence its subsequence converges to $\dot p(t)$ *almost everywhere* on $[a,b]$. Passing finally to the limit in (6.6) and taking into account the established *pointwise* convergence together with (6.5), we arrive at the extended Euler–Lagrange inclusion (6.3) and complete the proof of the theorem. □

Note that the results obtained in Theorem 6.1 are different from those derived in [14, subsection 6.1.5] not only because of the *absence* of any SNC-like assumptions on the target/constraint set but also because here the "coderivative normality" property is *not* imposed on F, as is needed in [14] in similar settings. Observe also that the arguments developed above allow us to provide the correspondent improvements in the case of Lipschitzian endpoint constraints of the Euler–Lagrange-type necessary optimality conditions derived in [15] for evolution models governed by *semilinear inclusions*

$$\dot x(t)\in Ax(t)+F\big(x(t),t\big),\tag{6.7}$$

where A is an *unbounded* infinitesimal generator of a *compact* C_0*-semigroup* on X, and where continuous solutions to (6.7) are understood in the *mild* sense.

7. Euler–Lagrange and maximum conditions with no relaxation. The main objective of this section is to derive necessary optimality conditions for intermediate local minimizers $\bar x(\cdot)$ of evolution inclusions *without any relaxation*. We show that it can be done under certain more restrictive assumptions on the initial data in

comparison with those in Theorem 6.1. For simplicity, we consider here the *Mayer version* (P_M) of problem (P) with $\vartheta = 0$ in (1.1). In this case, the *Euler–Lagrange inclusion* (6.3) admits the *coderivative form*

$$\dot{p}(t) \in \operatorname{clco} D^*F\big(\bar{x}(t), \dot{\bar{x}}(t)\big)\big(-p(t)\big) \ \text{ a.e. } \ t \in [a,b], \tag{7.1}$$

which easily implies, due to the extremal property for coderivatives of convex-valued mappings given in [13, Theorem 1.34], the *Weierstrass–Pontryagin maximum condition*

$$\big\langle p(t), \dot{\bar{x}}(t)\big\rangle = \max_{v \in F(\bar{x}(t))} \big\langle p(t), v\big\rangle \ \text{ a.e. } \ t \in [a,b], \tag{7.2}$$

provided that the sets $F(x)$ are *convex* near $\bar{x}(t)$ for a.e. $t \in [a,b]$. Our goal is to justify the above Euler–Lagrange and Weierstrass–Pontryagin conditions, together with the other necessary optimality conditions of Theorem 6.1, for intermediate minimizers of the Mayer problem (P_M) subject to the Lipschitzian endpoint constraints (1.3) and (1.4), *without any convexity or relaxation* assumptions and with *no SNC-like* requirements imposed on the endpoint constraint set. To accomplish this goal, we employ a certain approximation technique involving *Ekeland's variational principle* combined with other advanced results of variational analysis and generalized differentiation, which allow us to reduce the constrained problem under consideration to an unconstrained (and thus *stable with respect to relaxation*) Bolza problem studied in section 6. However, this requires additional assumptions on the initial data of (P_M) imposed in what follows.

Recall that a set-valued mapping $F\colon X \rightrightarrows Y$ is *strongly coderivatively normal* at $(\bar{x}, \bar{y}) \in \operatorname{gph} F$ if its normal coderivative (4.3) admits the representation

$$\begin{aligned} D^*F(\bar{x},\bar{y})(y^*) = \Big\{ x^* \in X^* \Big| \ &\exists \text{ sequences } \ (x_k, y_k) \to (\bar{x}, \bar{y}),\ x_k^* \overset{w^*}{\to} x^*, \ \text{ and } \ y_k^* \to y^* \\ &\text{with } \ y_k \in F(x_k) \ \text{ and } \ x_k^* \in \widehat{D}^*F(x_k, y_k)(y_k^*) \ \text{ as } \ k \to \infty \Big\} \\ =: D_M^*F(\bar{x},\bar{y})(y^*), \end{aligned} \tag{7.3}$$

where $D_M^*F(\bar{x},\bar{y})$ is called the *mixed coderivative* of F at $(\bar{x},\bar{y})$. Observe that the only difference between the normal and mixed coderivatives of F at $(\bar{x},\bar{y})$ is that the *mixed* weak* convergence of $x_k^* \overset{w^*}{\to} x^*$ and the norm convergence of $y_k^* \to y^*$ is used for $D_M^*F(\bar{x},\bar{y})$ in (7.3), in contrast to the weak* convergence of *both* components $(x_k^*, y_k^*) \overset{w^*}{\to} (x^*, y^*)$ for $D^*F(\bar{x},\bar{y})$ in (4.3) via (4.1). Besides the obvious case of $\dim Y < \infty$, the strong coderivative normality holds in many important infinite-dimensional settings, and the property is preserved under various compositions; see [13, Proposition 4.9] describing major classes of mappings satisfying this property.

A mapping $F\colon X \rightrightarrows Y$ is called *sequentially normally compact* (SNC) at $(\bar{x},\bar{y}) \in \operatorname{gph} F$ if for any sequences $(x_k, y_k) \overset{\operatorname{gph} F}{\to} (\bar{x},\bar{y})$ and $(x_k^*, y^*) \in \widehat{N}\big((x_k, y_k); \operatorname{gph} F\big)$ one has

$$(x_k^*, y_k^*) \overset{w^*}{\to} 0 \Longrightarrow \|(x_k^*, y_k^*)\| \to 0 \ \text{ as } \ k \to \infty.$$

As was discussed in section 1, this property is a far-reaching extension of the "finite-codimension" and other related properties of sets and mappings. It always holds in finite dimensions, while in reflexive spaces it agrees with the "compactly epi-Lipschitzian" property by Borwein and Strójwas; see [13] for more details, discussions, and calculus.

Finally, recall that the given norm on a Banach space X is *Kadec* if the strong and weak convergences agree on the boundary of the unit sphere of X. It is well known that every reflexive space admits an equivalent Kadec norm.

THEOREM 7.1 (Euler–Lagrange and Weierstrass–Pontryagin conditions for intermediate local minimizers with no relaxation). *Let $\bar{x}(\cdot)$ be an intermediate local minimizer for the Mayer problem (P_M) in* (1.1)–(1.4) *under the standing hypotheses* (H1) *and* (H3) *on F and φ_i. Assume in addition that*

(a) *the state space X is separable and reflexive with the Kadec norm on it;*

(b) *the velocity mapping F is SNC at $\big(\bar{x}(t),\dot{\bar{x}}(t)\big)$ and strongly coderivatively normal with weakly closed graph around this point for a.e. $t\in[a,b]$.*

Then there are nontrivial Lagrange multipliers $0\neq(\lambda_0,\ldots,\lambda_{m+r})\in\mathbb{R}^{m+r+1}$ and an absolutely continuous mapping $p\colon[a,b]\to X^*$ satisfying the following relationships:

- *the sign and complementarity slackness conditions in* (6.1) *and* (6.2)*;*
- *the Euler–Lagrange inclusion* (7.1)*, where the closure operation is redundant;*
- *the Weierstrass–Pontryagin maximum condition* (7.2)*; and*
- *the transversality inclusion* (6.4)*.*

Proof. Denote

$$\varphi_0^+(x,\nu):=\max\big\{\varphi_0(x)-\nu,0\big\},\quad \varphi_i^+(x):=\max\big\{\varphi_i(x),0\big\},\ i=1,\ldots,m, \tag{7.4}$$

and, following the *method of metric approximations* [11], consider the parametric cost functional

$$\theta_\nu[x]:=\left[\big(\varphi_0^+\big)^2\big(x(b),\nu\big)+\sum_{i=1}^m\big(\varphi_i^+\big)^2\big(x(b)\big)+\sum_{i=m+1}^{m+r}\varphi_i^2\big(x(b)\big)\right]^{1/2},\quad \nu\in\mathbb{R}, \tag{7.5}$$

over trajectories for the evolution inclusion (1.1) with *no endpoint constraints*. Since $\bar{x}(\cdot)$ is an *intermediate local minimizer* for (P_M) and due to the constructions in (7.4) and (7.5), we have

$$\theta_\nu[x]>0\ \text{ for any }\ \nu<\bar{\nu}:=\varphi_0\big(\bar{x}(b)\big), \tag{7.6}$$

provided that $x(\cdot)$ is a trajectory for (1.2) belonging to the prescribed $W^{1,1}$-neighborhood of the given intermediate local minimizer and such that $x(t)\in U$ for all $t\in[a,b]$, where the open set $U\subset X$ is taken from the requirements in (H1) imposed on $\bar{x}(\cdot)$.

Form as in [2] the space $\mathcal{X}$ of all the trajectories $x(\cdot)$ for (1.2) satisfying *the only constraint* $x(t)\in\operatorname{cl}U$ as $t\in[a,b]$ with the metric

$$d[x,y]:=\int_a^b\|\dot{x}(t)-\dot{y}(t)\|\,dt.$$

We can easily check, based on Definition 2.1 of solutions to the original differential inclusion and on standard properties of the Bochner integral, that the metric space $\mathcal{X}$ is *complete* and that the function $\theta_\nu[\cdot]$ is (Lipschitz) continuous on $\mathcal{X}$ for any $\nu\in\mathbb{R}$. It follows from the above constructions that for every $\varepsilon>0$ there is $\nu_\varepsilon<\bar{\nu}$ such that $\nu_\varepsilon\to\bar{\nu}$ as $\varepsilon\downarrow0$ and

$$0\le\theta_\varepsilon[\bar{x}]<\varepsilon\le\inf_{x\in\mathcal{X}}\theta_\varepsilon[x]+\varepsilon\quad\text{with}\ \ \theta_\varepsilon:=\theta_{\nu_\varepsilon}.$$

Now applying the classical *Ekeland variational principle*, find an arc $x_\varepsilon(\cdot)\in\mathcal{X}$ satisfying

$$d[x_\varepsilon,\bar{x}]\le\sqrt{\varepsilon}\ \text{ and }\ \theta_\varepsilon[x]+\sqrt{\varepsilon}d[x,x_\varepsilon]\ge\theta_\varepsilon[x_\varepsilon]\ \text{ for all }\ x\in\mathcal{X}. \tag{7.7}$$

This distance estimate yields that $x_\varepsilon(t) \in U$ as $t \in [a,b]$ and that $x_\varepsilon(\cdot)$ belongs to the fixed $W^{1,1}$-neighborhood of the intermediate local minimizer $\bar{x}(\cdot)$ whenever $\varepsilon > 0$ is sufficiently small. Hence $\theta_\varepsilon[x_\varepsilon] > 0$ for such ε by (7.6).

Given positive numbers ε and $\eta > 0$, we form the Bolza functional

$$J_{\varepsilon,\eta}[x] := \theta_\varepsilon[x] + \sqrt{\varepsilon} d[x, x_\varepsilon] + \eta\sqrt{1+\ell_F^2} \int_a^b \operatorname{dist}\big((x(t), \dot{x}(t)); \operatorname{gph} F\big)\, dt$$

and show, following the proof of the claim in [14, Theorem 6.27], that there is a number $\eta \geq 1$ such that for every $\varepsilon \in (0, 1/\eta)$ the arc $x_\varepsilon(\cdot)$ built above is an *intermediate local minimizer* for this functional over all absolutely continuous mappings $x\colon [a,b] \to X$, *not necessarily trajectories* for (1.1), satisfying the constraints

$$x(a) = x_0 \;\text{ and }\; x(t) \in U \;\text{ for }\; t \in [a,b],$$

where the one $x(t) \in U$ can be *ignored* from the viewpoint of necessary optimality conditions, since the set U is open in X. Taking into account the structures of $\theta_\varepsilon[\cdot]$ and $d[\cdot,\cdot]$, we conclude that $x_\varepsilon(\cdot)$ is an intermediate minimizer for the following *unconstrained* Bolza problem with *Lipschitzian* data:

$$\text{minimize } \; \varphi_\varepsilon\big(x(b)\big) + \int_a^b \vartheta_\varepsilon\big(x(t), \dot{x}(t), t\big)\, dt \tag{7.8}$$

over absolutely continuous arcs $x\colon [a,b] \to X$ satisfying $x(a) = x_0$ and lying in a $W^{1,1}$-neighborhood of $\bar{x}(\cdot)$, where the functions $\varphi_\varepsilon\colon X \to \mathbb{R}$ and $\theta_\varepsilon\colon X \times X \times [a,b] \to \mathbb{R}$ are given by

$$\varphi_\varepsilon(x) := \left[(\varphi_0^+)^2(x, \nu_\varepsilon) + \sum_{i=1}^m \big(\varphi_i^+\big)^2(x) + \sum_{i=m+1}^{m+r} \varphi_i^2(x) \right]^{1/2}, \tag{7.9}$$

$$\vartheta_\varepsilon(x,v,t) := \eta\sqrt{1+\ell_F^2}\, \operatorname{dist}\big((x,v); \operatorname{gph} F\big) + \sqrt{\varepsilon}\|v - \dot{x}_\varepsilon(t)\|. \tag{7.10}$$

To apply the results of Theorem 6.1 to the case of problem (7.8), we first note that *every intermediate local minimizer* for the unconstrained problem (7.8) provides a *relaxed* intermediate local minimum for this problem. It follows from the *relaxation stability* of unconstrained Bolza problems with finite integrands, which is ensured by an appropriate infinite-dimensional extension of the classical Bogolyubov theorem valid under the assumptions made; see Lemma 2.3 above and its "intermediate" local counterpart given in [9, Theorem 4] whose proof holds in the infinite-dimensional setting under consideration. Furthermore, observe that although Theorem 6.1 is presented for autonomous problems, its results hold true with no change for the case of *summable* integrands as in (7.10); it can be justified similarly to the proof of [14, Theorem 6.22] given for problems with geometric endpoint constraints. Finally, it follows from the proof of Theorem 6.1 that the compactness of the velocity sets assumed in (H1) is, in fact, *not needed* for the unconstrained and $W^{1,1}$-bounded framework of the Bolza problem (7.8).

Applying the optimality conditions of Theorem 6.1 to problem (7.8) with the initial data (7.9) and (7.10), for all small $\varepsilon > 0$, we find an absolutely continuous adjoint arc $p_\varepsilon\colon [a,b] \to X^*$ satisfying

$$\dot{p}_\varepsilon(t) \in \operatorname{co}\Big\{ u \in X^* \;\Big|\; \big(u, p_\varepsilon(t)\big) \in \mu\, \partial \operatorname{dist}\big((x_\varepsilon(t), \dot{x}_\varepsilon(t)); \operatorname{gph} F\big) + \sqrt{\varepsilon}\big(0, \mathbb{B}^*\big) \Big\} \tag{7.11}$$

for a.e. $t\in[a,b]$ with $\mu:=\eta\sqrt{1+\ell_F^2}$ and

$$-p_\varepsilon(b)\in\partial\left[(\varphi_0^+)^2(\cdot,\nu_\varepsilon)+\sum_{i=1}^m\big(\varphi_i^+\big)^2(\cdot)+\sum_{i=m+1}^{m+r}\varphi_i^2(\cdot)\right]^{1/2}\big(x_\varepsilon(b)\big). \tag{7.12}$$

Note that the last term on the right-hand side of (7.11) appears due to applying the sum rule from [13, Theorem 2.33(c)] to the integrand (7.10) and using the well-known subdifferential formula for the norm function. The other difference between (7.10) and (6.3) is that (7.11) *does not contain the closure operation* as in (6.3). The norm-closure operation can be omitted in (7.11), since the basic subdifferential sets for Lipschitzian functions are weak compact in reflexive spaces (which are weakly compactly generated) by [13, Theorem 3.59(i)], and hence the right-hand side of (7.11) is closed in the norm topology of the dual space X^*.

To deal further with (7.11), fix $t\in[a,b]$ and consider the two possible cases for the location of $\big(x_\varepsilon(t),\dot x_\varepsilon(t)\big)$ relative to the graph of the velocity mapping $F(\cdot)$:

$$\text{(i) }\big(x_\varepsilon(t),\dot x_\varepsilon(t)\big)\in\operatorname{gph}F\quad\text{ and }\quad\text{(ii) }\big(x_\varepsilon(t),\dot x_\varepsilon(t)\big)\notin\operatorname{gph}F.$$

In case (i) we use [13, Theorem 1.97] on basic subgradients of the distance function at set points, which gives the *approximate adjoint inclusion*

$$\dot p_\varepsilon(t)\in\operatorname{co}\Big\{u\in X^*\Big|\ \big(u,p_\varepsilon(t)\big)\in N\big((x_\varepsilon(t),\dot x_\varepsilon(t));\operatorname{gph}F\big)+\sqrt{\varepsilon}\big(0,\mathbb{B}^*\big)\Big\}. \tag{7.13}$$

The out-of-set case (ii) is more involved and requires the *Kadec norm* structure of X together with the *weak closedness* assumption on the graph of F. Then we obtain, by [13, Theorem 1.105], the relationship

$$\partial\operatorname{dist}\big((x_\varepsilon(t),\dot x_\varepsilon(t));\operatorname{gph}F\big)\subset\bigcup_{(x,v)\in\Pi\left((x_\varepsilon(t),\dot x_\varepsilon(t));\,\operatorname{gph}F\right)}N\big((x,v);\operatorname{gph}F\big)$$

via the *projection operator* $\Pi(\cdot;\operatorname{gph}F)$ at the reference point. Taking into account the a.e. pointwise convergence $\big(x_\varepsilon(t),\dot x_\varepsilon(t)\big)\to\big(\bar x(t),\dot{\bar x}(t)\big)$ as $\varepsilon\downarrow0$ that follows from (7.7), we come up to a modified inclusion (7.13) with the replacement of $\big(x_\varepsilon(t),\dot x_\varepsilon(t)\big)$ by some sequence $(\widetilde x_\varepsilon,\widetilde v_\varepsilon)\overset{\operatorname{gph}F}{\to}\big(\bar x(t),\dot{\bar x}(t)\big)$ as $\varepsilon\downarrow0$, while we keep the form (7.13) for simplicity.

Consider next the transversality condition (7.12) with φ_i^+ defined in (7.4). Employing the sum and chain rules [13, subsection 3.2.1] for basic subgradients in (7.12) and taking into account relationships (7.5) and (7.6) with $\nu_\varepsilon\uparrow\bar\nu$ as $\varepsilon\downarrow0$, we have

$$-p_\varepsilon(b)\in\sum_{i=0}^m\lambda_{i\varepsilon}\partial\varphi_i\big(x_\varepsilon(b)\big)+\sum_{i=m+1}^{m+r}\lambda_{i\varepsilon}\Big[\partial\varphi_i\big(x_\varepsilon(b)\big)\bigcup\partial\big(-\varphi_i\big)\big(x_\varepsilon(b)\big)\Big], \tag{7.14}$$

where the multipliers $\lambda_{i\varepsilon}$ satisfy the conditions

$$\lambda_{i\varepsilon}\ge0\ \text{ for all }\ i=0,\dots,m+r,\quad\sum_{i=0}^{m+r}\lambda_{i\varepsilon}^2=1\ \text{ as }\ \varepsilon\downarrow0. \tag{7.15}$$

By (7.15), we suppose without loss of generality that $\lambda_{i\varepsilon}\to\lambda_i$ as $\varepsilon\downarrow0$ for $i=0,\dots,m+r$, where the limiting multipliers λ_i are not zero simultaneously and satisfy

the sign and complementary slackness conditions in (6.1) and (6.2). Furthermore, it follows from (7.14) and (7.15) that the family $\{p_\varepsilon(b)\}_{\varepsilon>0}$ is uniformly bounded in X^* for ε sufficiently small. Proceeding as in the proof of Theorem 6.1, we observe that the strong coderivative normality assumption on F allows us, by (7.13), to use the *mixed coderivative characterization* of the Lipschitz property of F from [13, Corollary 4.11] and thus to find an absolutely continuous arc $p\colon[a,b]\to X^*$ such that $\dot p_\varepsilon(\cdot)\to\dot p(\cdot)$ weakly in $L^1\big([a,b];X^*\big)$ and $p_\varepsilon(t)\to p(t)$ weakly in X^* as $\varepsilon\downarrow 0$ for each $t\in[a,b]$.

To complete the proof of the Euler–Lagrange and transversality inclusions of the theorem, we pass to the limit in (7.13) and (7.14) as $\varepsilon\downarrow 0$ by using the Mazur theorem on the strong convergence of convex combinations for $\{\dot p_\varepsilon(\cdot)\}$. To accomplish this limiting procedure and to arrive at the desired inclusions (7.1) and (6.4), we use the *closed-graph* property of the basic normal cone in (7.13) and the basic subdifferential in (7.14). This follows from [13, Theorem 3.60] due to the SNC assumption on F and the Lipschitz continuity of φ_i in the reflexive state space X. Observe that the closedness operation in (7.1) is redundant, similar to (7.13), due to the *uniform boundedness* of $\big\{\big(\dot p_\varepsilon(\cdot),p_\varepsilon(\cdot)\big)\big\}$ in $X^*\times X^*$ and the arguments above involving now [13, Theorem 3.59(ii)].

The given proof justifies the extended Euler–Lagrange and transversality conditions in the theorem for arbitrary intermediate local minimizers to problem (P_M) with *no relaxation*. In the general nonconvex setting the Euler–Lagrange inclusion (7.1) does not automatically imply the maximum condition (7.2). To establish the latter condition supplementing the other necessary conditions of the theorem, we follow the proof of [19, Theorem 7.4.1] given for a Mayer problem of type (P_M) involving nonconvex differential inclusions in finite-dimensional spaces; it holds with minor changes in infinite dimensions under the assumptions imposed. The proof of the latter theorem is based on reducing the constrained Mayer problem for nonconvex differential inclusions to an unconstrained Bolza (finite Lagrangian) problem, which in turn is reduced to a problem of optimal control with *smooth dynamics* and *nonsmooth endpoint constraints* first treated in [11] via the nonconvex normal cone (4.1) and the corresponding subdifferential (4.5) introduced therein to describe the appropriate transversality conditions in the maximum principle. □

REFERENCES

[1] J. M. Borwein and Q. J. Zhu, *Techniques of Variational Analysis*, CMS Books Math., Springer, New York, 2005.

[2] F. H. Clarke, *Necessary conditions for a general control problem*, in Calculus of Variations and Control Theory, D. L. Russel, ed., Academic Press, New York, 1976, pp. 257–278.

[3] F. H. Clarke, *Necessary conditions in dynamic optimization*, Mem. Amer. Math. Soc., 173 (2005), no. 816.

[4] J. Diestel and J. J. Uhl, Jr., *Vector Measures*, AMS, Providence, RI, 1977.

[5] F. S. De Blasi, G. Pianigiani, and A. A. Tolstonogov, *A Bogolyubov-type theorem with a nonconvex constraint in Banach spaces*, SIAM J. Control Optim., 43 (2004), pp. 466–476.

[6] K. Deimling, *Multivalued Differential Equations*, De Gruyter, Berlin, 1992.

[7] H. O. Fattorini, *Infinite Dimensional Optimization and Control Theory*, Cambridge University Press, Cambridge, UK, 1999.

[8] A. D. Ioffe, *Euler-Lagrange and Hamiltonian formalisms in dynamic optimization*, Trans. Amer. Math. Soc., 349 (1997), pp. 2871–2900.

[9] A. D. Ioffe and R. T. Rockafellar, *The Euler and Weierstrass conditions for nonsmooth variational problems*, Calc. Var. Partial Differential Equations, 4 (1996), pp. 59–87.

[10] X. Li and J. Yong, *Optimal Control Theory for Infinite-Dimensional Systems*, Birkhäuser, Boston, 1995.

[11] B. S. Mordukhovich, *Maximum principle in problems of time optimal control with nonsmooth*

constraints, J. Appl. Math. Mech., 40 (1976), pp. 960–969.
[12] B. S. Mordukhovich, *Discrete approximations and refined Euler–Lagrange conditions for nonconvex differential inclusions*, SIAM J. Control Optim., 33 (1995), pp. 882–915.
[13] B. S. Mordukhovich, *Variational Analysis and Generalized Differentiation,* I: *Basic Theory*, Grundlehren Math. Wiss. 330, Springer, Berlin, 2006.
[14] B. S. Mordukhovich, *Variational Analysis and Generalized Differentiation,* II: *Applications*, Grundlehren Math. Wiss. 331, Springer, Berlin, 2006.
[15] B. S. Mordukhovich and D. Wang, *Optimal control of semilinear evolution inclusions via discrete approximations*, Control Cybernet., 34 (2005), pp. 849–870.
[16] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Grundlehren Math. Wiss. 317, Springer, Berlin, 1998.
[17] G. V. Smirnov, *Introduction to the Theory of Differential Inclusions*, AMS, Providence, RI, 2001.
[18] A. A. Tolstonogov, *Differential Inclusions in a Banach Space*, Kluwer, Dordrecht, The Netherlands, 2000.
[19] R. B. Vinter, *Optimal Control*, Birkhäuser, Boston, 2000.
[20] R. B. Vinter and P. D. Woodford, *On the occurrence of intermediate local minimizers that are not strong local minimizers*, Systems Control Lett., 31 (1997), pp. 235–242.

SIAM J. OPTIM.
Vol. 18, No. 3, pp. 778–788

TWO-STAGE STOCHASTIC PROGRAMS WITH MIXED PROBABILITIES*

PAUL BOSCH†, ALEJANDRO JOFRÉ‡, AND RÜDIGER SCHULTZ§

Abstract. We extend the traditional two-stage linear stochastic program by probabilistic constraints imposed in the second stage. This adds nonlinearity such that basic arguments for analyzing the structure of linear two-stage stochastic programs have to be rethought from the very beginning. We identify assumptions under which the problem is structurally sound and behaves stably under perturbations of probability measures.

Key words. stochastic programming, two-stage models, stability analysis

AMS subject classifications. 90C15, 90C30

DOI. 10.1137/050648754

1. Introduction.

Stochastic programming deals with the optimization of decision making under uncertainty over time. When building stochastic programming models, two main approaches consist in introducing future costs and in fixing certain reliability levels for constraints. For earlier reviews on the various aspects in stochastic programming we refer, for example, to [5], [17], [22], [23], and [24].

Traditional two-stage stochastic programming is concerned with problems that require a here-and-now decision on the basis of given probabilistic information on the random data without making further observations. The costs to be minimized consist of the direct costs of the first-stage decision as well as the costs generated by the need of taking a recourse (or second-stage) decision in response to the random environment. The stability behavior of stochastic programming models when changing the probability measure is studied in many papers. We refer to the qualitative studies (continuity properties of optimal values and optimal solutions) found in [1], [7], [12], and [23] and to work on quantitative stability (quantitative continuity of optimal values and optimal solutions) found in [6], [14], [19], [20], and [21]. The bibliographical notes in section 5 of [14] provide a detailed account of the developments in stability analysis of stochastic programs.

Motivated by the study of stochastic programming problems coming from planning and operational management in power generation companies, we introduce the following parametric family of mixed-probability stochastic programs:

$$P(\mu,\lambda) \qquad \min\left\{c^T x + \int_{\mathbb{R}^s} Q(z-Ax,\lambda)\,\mu(dz) \;:\; x\in C\right\}, \quad (\mu,\lambda)\in\Delta\times\Lambda,$$

where

$$Q(t,\lambda) \;:=\; \min\{q^T y \;:\; Wy=t,\, y\geq 0,\, \lambda(H_j(y))\geq p_j, j=1,\dots,d\}. \tag{1.1}$$

*Received by the editors December 31, 2005; accepted for publication (in revised form) November 15, 2006; published electronically October 4, 2007.
http://www.siam.org/journals/siopt/18-3/64875.html

†Facultad de Ingeniería, Universidad Diego Portales, Ave. Ejército 441, 5to. piso, Santiago, Chile (paul.bosch@prof.udp.cl).

‡Centro de Modelamiento Matemático and Departamento Ingenieria Matematica, Universidad de Chile, Ave. Blanco Encalada, Santiago, Chile (ajofre@dim.uchile.cl).

§Department of Mathematics, University of Duisburg–Essen, Lotharstr. 65, D-47048 Duisburg, Germany (schultz@math.uni-duisburg.de).

Here $H_j, j = 1, \dots, d$, are set-valued mappings from $\mathbb{R}^m$ to $\mathbb{R}^r$ with closed graph, and $p_j, j = 1, \dots, d$, are predesigned probability levels. If $\mathcal{P}(\mathbb{R}^s)$, $\mathcal{P}(\mathbb{R}^r)$ denote the sets of all Borel probability measures on $\mathbb{R}^s$ and $\mathbb{R}^r$, respectively, we assume that Δ, Λ are subsets of $\mathcal{P}(\mathbb{R}^s)$, $\mathcal{P}(\mathbb{R}^r)$. Moreover, C is a closed subset of $\mathbb{R}^{\bar{m}}$, and all remaining vectors and matrices have suitable dimensions.

In the case of planning and operational management in power generation companies, the first stage variable x in the model represents generation capacity investment decisions, such as changes (continuous) of maximum generation capacity for thermal plants. The variable z is a random demand and y is the second-stage operational variable representing the level of production of energy. The latter is also limited by emission rights for carbon dioxide that may concern single plants or consortia of plants. The level of permitted emission is considered random, since emission rights are traded at predesigned markets via auctions, for instance, whose outcomes are uncertain to market participants. This motivates the modeling of limitations on the operational variables resulting from emission rights by probabilistic rather than deterministic constraints.

As an illustrative example consider a power system consisting of plants $i = 1, \dots, I$ to be operated over a time horizon with subintervals $t = 1, \dots, T$. The (first-stage) variables x_i denote (continuous) capacity expansions for the individual plants. They must be within the bounds

$$0 \ \leq \ x_i \ \leq \ C_i, \qquad i = 1, \dots, I, \tag{1.2}$$

and they incur the costs

$$\sum_{i=1}^{I} c_i x_i. \tag{1.3}$$

The (second-stage) variable y_{it} denotes the power output of plant i at time t. Constraints concern the coverage of power demand over the whole planning horizon,

$$\sum_{i=1}^{I} y_{it} \ \geq \ z_t(\omega), \qquad t = 1, \dots, T, \tag{1.4}$$

and limitation of individual outputs by the base capacity b_i plus the capacity expansion x_i:

$$0 \ \leq \ y_{it} \ \leq \ b_i + x_i, \qquad i = 1, \dots, I, \ t = 1, \dots, T. \tag{1.5}$$

Production costs are assumed to be linear and sum up to

$$\sum_{t=1}^{T} \sum_{i=1}^{I} q_i y_{it}. \tag{1.6}$$

A generic formulation for the probabilistic limitations on the second-stage variables induced by emission rights reads as follows:

$$\lambda \left(\left\{ \bar{a}_j \ : \ \sum_{t \in \mathcal{T}_j} \sum_{i \in \mathcal{I}_j} \beta_i y_{it} \ \leq \ \bar{a}_j \right\} \right) \ \geq \ p_j, \qquad j = 1, \dots, d. \tag{1.7}$$

Here, $\mathcal{T}_j, \mathcal{I}_j$ are subsets of $\{1,\dots,T\}$ and $\{1,\dots,I\}$, respectively, to express limitations summed up over subhorizons and clusters of plants. The coefficients β_i denote emission intensities that may vary from plant to plant. The random levels of permitted emissions for the different subhorizons and clusters are the $\bar{a}_j = \bar{a}_j(\omega)$. Finally, p_j are the probability levels with which the limitations are to be met.

The model (1.2)–(1.7) illustrates in a simple example the relevance of the general model $P(\mu,\lambda)$. Further operational details of power planning such as ramp rates for power output in adjacent time intervals or fuel constraints can easily be added without leaving the frame of $P(\mu,\lambda)$. A next step is to go beyond $P(\mu,\lambda)$ by including unit commitment features and, thus, integer variables; see [16] for a first attempt in that direction, including numerical experiments.

In our example, the chance constraints (1.7) mathematically really matter for the second-stage optimization problem (1.1). Without them, the remaining second-stage problem, given by (1.4)–(1.6), is very simple. It is separable in t, and the individual subproblems are continuous knapsack problems that solve immediately.

Coming back to $P(\mu,\lambda)$ we see that this model extends the traditional two-stage linear stochastic program by introducing some probabilistic constraints $\lambda(H_j(y)) \geq p_j, j = 1,\dots,d$, in the second stage of the problem. These types of constraints add nonlinearities to the problem so that basic arguments to analyze the well-posedness of $P(\mu,\lambda)$ have to be rethought from the very beginning. The purpose of this paper is to identify assumptions under which $P(\mu,\lambda)$ is structurally well-behaved and stable under perturbations of (μ,λ).

Somewhat related two-stage stochastic programming formulations with probabilistic constraints are presented in Chapter 12.10 of [11]. In a first group of models, probabilistic constraints serve to replace the induced constraints by the requirement that the second-stage problem is solvable with a prescribed probability. Another group of two-stage models has probabilistic constraints in both stages. Neither model type had been studied any further at the time of the writing of [11], nor, to the best of our knowledge, were they further studied since then.

2. Basic well-posedness.

LEMMA 1. *For fixed $\lambda \in \Lambda$ and $t \in \mathbb{R}^s$ the constraint set in* (1.1) *is closed.*

Proof. Since the graphs of $H_j, j = 1,\dots,d$, are all closed we have

$$\limsup_{y_n \to y_o} H_j(y_n) \;\subseteq\; H_j(y_o), \quad j = 1,\dots,d,$$

where $\limsup_{y_n \to y_o} H_j(y_n)$ denotes the set of all points belonging to infinitely many of the sets $H_j(y_n)$. Hence, by the semicontinuity of the probability measure (see, for example, [4]) we have

$$\lambda(H_j(y_o)) \geq \lambda\left(\limsup_{y_n \to y_o} H_j(y_n)\right) \geq \limsup_{y_n \to y_o} \lambda(H_j(y_n)) \geq p_j$$

for all $j = 1,\dots,d$ and all sequences $y_n \to y_o$, which allows us to conclude the closedness of the constraint set (1.1). □

The major difficulty in understanding the structure of $P(\mu,\lambda)$ rests in a dilemma about the function Q. On the one hand, Q is the optimal-value function of a nonlinear program with parameters t and λ. Parametric optimization mainly provides local results about the structure of Q. Global results are very scarce and require specific assumptions that are often hard to verify. On the other hand, Q arises as an integrand in $P(\mu,\lambda)$. For studying properties of the related integral, global information about

Q is required. From this viewpoint, it is not surprising that most of the structural results about two-stage stochastic programs concern the purely linear and the linear mixed-integer cases, i.e., the widest problem classes where parametric optimization offers broader results about global stability.

To lay a foundation for the structural analysis of Q we formulate the following general assumptions:

(A1) For any $\lambda \in \Lambda$ there exist a nonempty set $\mathcal{R}_\lambda \subseteq \mathbb{R}^s$ and a Lebesgue null set $\mathcal{N}_\lambda \subseteq \mathbb{R}^s$ such that the function $Q(\cdot, \lambda)$ is real-valued and measurable on $\mathcal{R}_\lambda$ and continuous on $\mathcal{R}_\lambda \setminus \mathcal{N}_\lambda$.

(A2) It holds that

$$\bigcup_{\mu \in \Delta} \text{supp } \mu \ \subseteq \ \bigcap_{\lambda \in \Lambda} \bigcap_{x \in C} \{Ax + \mathcal{R}_\lambda\},$$

where supp μ denotes the smallest closed set in $\mathbb{R}^s$ with μ-measure 1.

(A3) There exists a real-valued, measurable function h on $\mathbb{R}^s$, which we call the bounding function, with the following properties:

1. (Q-majorization.) It holds that $|Q(t, \lambda)| \leq h(t)$ for all $t \in \mathcal{R}_\lambda$ and all $\lambda \in \Lambda$.
2. (Integrability.) It holds that $\int_{\mathbb{R}^s} h(z)\, \mu(dz) < +\infty$ for all $\mu \in \Delta$.
3. (Generalized subadditivity.) There exists a $\kappa > 0$ such that $h(t_1 + t_2) \leq \kappa(h(t_1) + h(t_2))$ for all $t_1, t_2 \in \mathbb{R}^s$.
4. (Local boundedness.) For each $t \in \mathbb{R}^s$ there exists an open neighborhood of t where h is bounded.

LEMMA 2. *Assume* (A1), (A2), *and* (A3). *Then*

$$G(x, \mu, \lambda) \ := \ \int_{\mathbb{R}^s} Q(z - Ax, \lambda)\, \mu(dz) \tag{2.1}$$

is real-valued for all $x \in C, \lambda \in \Lambda, \mu \in \Delta$.

Proof. Let $x \in C, \lambda \in \Lambda, \mu \in \Delta$ be fixed. By (A2) we have $z - Ax \in \mathcal{R}_\lambda$ for all $z \in$ supp μ. Together with (A1) this yields that $Q(\cdot - Ax, \lambda)$ is measurable on supp μ. By (A3.1) and (A3.3) it holds that

$$|Q(z - Ax, \lambda)| \ \leq \ h(z - Ax) \ \leq \ \kappa h(z) + \kappa h(-Ax)$$

for all $z \in$ supp μ. By (A3.2) the right-hand side above is μ-integrable, and the assertion follows. □

LEMMA 3. *Assume* (A1), (A3), *and*

(A2*) *there exists an open set* $C^* \supseteq C$ *such that* $\bigcup_{\mu \in \Delta} \text{supp } \mu \subseteq \bigcap_{\lambda \in \Lambda} \bigcap_{x \in C^*} \{Ax + \mathcal{R}_\lambda\}$.

Let $x \in C, \lambda \in \Lambda, \mu \in \Delta$ *such that* $\mu(Ax + \mathcal{N}_\lambda) = 0$. *Then* $G(\cdot, \mu, \lambda)$ *is continuous at* x.

Proof. Fix some (x, μ, λ) fulfilling the assumptions. Let $x_n \to x$ and assume without loss of generality that $x_n \in C^*$ for all n. (A2*) then implies $z - Ax_n \in \mathcal{R}_\lambda$ for all $z \in$ supp μ, and together with (A3.1)–(A3.3) it follows as in the proof of Lemma 2 that $G(x_n, \mu, \lambda) \in \mathbb{R}$ for all n. By (A1) and $\mu(Ax + \mathcal{N}_\lambda) = 0$ it follows that $Q(z - Ax_n, \lambda) \to Q(z - Ax, \lambda)$ for μ-almost all $z \in$ supp μ. By (A3.1) and (A3.3) it holds for all $z \in$ supp μ that

$$|Q(z - Ax_n, \lambda)| \ \leq \ h(z - Ax_n) \ \leq \ \kappa h(z) + \kappa h(-Ax_n),$$

yielding, together with (A3.2) and (A3.4), that there exists a uniform μ-integrable majorant for the functions $Q(\cdot - Ax_n)$. The assertion now follows from Lebesgue's theorem on dominated convergence. □

Remark 4. If μ is absolutely continuous with respect to the Lebesgue measure or, equivalently, has a density, then $\mu(Ax + \mathcal{N}_\lambda) = 0$ holds for all $x \in C$, since, by (A1), $Ax + \mathcal{N}_\lambda$ is a Lebesgue null set.

Remark 5. The essence of assumptions (A1)–(A3) is the following: Since $Q(\cdot, \lambda)$ is the optimal-value function of a minimization problem it well may attain the values $+\infty$ if the problem is infeasible and $-\infty$ if the problem is unbounded. Indeed, the following hold:

- (A1) makes sure that $Q(\cdot, \lambda)$ is finite on some set $\mathcal{R}_\lambda$.
- (A2) guarantees that the arguments $z - Ax$ are in $\mathcal{R}_\lambda$ for all relevant z and x. Otherwise, $Q(z - Ax, \lambda)$ would attain infinite values with positive probability, immediately preventing finiteness of the integral in (2.1).
- The continuity part of (A1) together with (A3) provides a framework for applying dominated convergence to show continuity of $G(\cdot, \mu, \lambda)$. Introducing the exceptional set $\mathcal{N}_\lambda$ in (A1) makes sense, since $Q(\cdot, \lambda)$ often lacks continuity on lower-dimensional subsets of its domain of finiteness. Furthermore, (A3) ensures an integrable upper-bound for the functions $|Q(\cdot - Ax, \lambda)|$ when x is varying in some neighborhood. Any other set of conditions ensuring this could be placed instead. Clearly, h reflects the global growth of $|Q(\cdot, \lambda)|$ whose quantitative analysis is acknowledged nontrivial for nonlinear problems.

Remark 6. Traditional two-stage linear stochastic programs form a class where verification of (A1)–(A3) is possible with low effort. There λ does not occur, and $Q(\cdot, \lambda)$ is given as

$$Q(t) := \min\{q^T y \,:\, Wy = t, y \geq 0\}.$$

With the assumptions of complete recourse ($W(\mathbb{R}^m_+) = \mathbb{R}^s$) and dual feasibility

$$\{u \in \mathbb{R}^s \,:\, W^T u \leq q\} \neq \emptyset$$

it holds that $Q(t) = \max_{k=1,\ldots,K} d_k^T t$, where d_k, $k = 1, \ldots, K$, are the vertices of $\{u \in \mathbb{R}^s \,:\, W^T u \leq q\}$. Then (A1) holds with $\mathcal{R}_\lambda = \mathcal{R} = \mathbb{R}^s, \mathcal{N}_\lambda = \mathcal{N} = \emptyset$, and (A2) becomes vacuous. The bounding function h in (A3) can be selected as $\kappa\|t\|$ with some positive constant κ. This reduces the integrability requirement to finiteness of the first moment of μ. For two-stage linear stochastic programs with integrality requirements in the second stage, similar results are valid, with the difference that the exceptional set $\mathcal{N}$ becomes effective, since continuity defects of Q may occur on lower-dimensional subsets [10].

The next proposition provides a sufficient condition for continuity of G jointly in the first two arguments. Convergence of the second argument is understood as weak convergence of probability measures on $\mathcal{P}(\mathbb{R}^s)$. We say that a sequence $\{\mu_n\}$ in $\mathcal{P}(\mathbb{R}^s)$ converges weakly to $\mu \in \mathcal{P}(\mathbb{R}^s)$, written $\mu_n \xrightarrow{w} \mu$, if for any bounded continuous function $g : \mathbb{R}^s \to \mathbb{R}$ it holds that

$$\int_{\mathbb{R}^s} g(z)\mu_n(dz) \to \int_{\mathbb{R}^s} g(z)\mu(dz) \quad \text{as} \quad n \to \infty.$$

A basic reference for weak convergence of probability measures is Billingsley's book [3].

PROPOSITION 7. *Assume* (A1), (A2*), (A3) *and that A has full row rank. Let $\rho > 1, \bar{h} > 0$ be constants and define*

$$\Delta_\rho \;:=\; \left\{\nu \in \Delta \;:\; \int_{\mathbb{R}^s} h(z)^\rho \,\nu(dz) \le \bar{h}\right\},$$

where h is the bounding function from (A3). *Let $x \in C, \lambda \in \Lambda, \mu \in \Delta_\rho$ such that $\mu(Ax + \mathcal{N}_\lambda) = 0$. Then $G(\cdot,\cdot,\lambda) : C^* \times \Delta_\rho \longrightarrow \mathbb{R}$ is continuous at (x,μ).*

Proof. Fix some (x,μ,λ) fulfilling the assumptions, and let $x_n \to x$, $\mu_n \xrightarrow{w} \mu$, $\mu_n \in \Delta_\rho$. Without loss of generality we assume that $x_n \in C^*$ for all n. Let $\mathcal{M} := \cup_{\nu\in\Delta_\rho} \operatorname{supp}\nu$ and define functions $h_n, h_o : \mathcal{M} \to \mathbb{R}$ by $h_n(z) := Q(z - Ax_n, \lambda)$ and $h_o(z) := Q(z - Ax, \lambda)$. Consider the set

$$E(x) \;:=\; \{z \in \mathcal{M} \;:\; \exists z_n \to z \text{ with } h_n(z_n) \not\to h_o(z)\}.$$

To see that $E(x) \subseteq Ax + \mathcal{N}_\lambda$ assume that $z \notin Ax + \mathcal{N}_\lambda$; hence $z - Ax \in \mathcal{R}_\lambda \setminus \mathcal{N}_\lambda$. By (A2*), there exists a neighborhood $N(x)$ of x such that $z - A\big(N(x)\big) \subseteq \mathcal{R}_\lambda$. Let $z_n \to z$ and rewrite $z_n - Ax_n = z - (Ax + z - z_n + Ax_n - Ax)$. Since A has full row rank, it holds that $Ax + z - z_n + Ax_n - Ax \in A\big(N(x)\big)$ for sufficiently large n. Therefore, $z_n - Ax_n \in \mathcal{R}_\lambda$ for sufficiently large n. Thus $h_n(z_n)$ is well-defined for these n. In view of (A1) we have $Q(z_n - Ax_n, \lambda) \to Q(z - Ax, \lambda)$; in other words, $h_n(z_n) \to h_o(z)$. Hence, $z \notin E(x)$.

By $\mu(Ax + \mathcal{N}_\lambda) = 0$ it follows that $\mu(E(x)) = 0$. Now Rubin's theorem on convergence of image measures (cf. [3]) yields

$$\mu_n \circ h_n^{-1} \xrightarrow{w} \mu \circ h_o^{-1}.$$

Next we establish the following uniform integrability

$$\lim_{a\to\infty} \sup_n \int_{|h_n(z)|\ge a} |h_n(z)|\,\mu_n(dz) \;=\; 0. \tag{2.2}$$

With $\rho > 1$ as in the assumptions we have

$$\begin{aligned}\int_{\mathcal{M}} |h_n(z)|^\rho \,\mu_n(dz) &\ge \int_{|h_n(z)|\ge a} |h_n(z)| \cdot |h_n(z)|^{\rho-1} \,\mu_n(dz)\\ &\ge a^{\rho-1} \cdot \int_{|h_n(z)|\ge a} |h_n(z)|\,\mu_n(dz).\end{aligned}$$

Therefore

$$\begin{aligned}\int_{|h_n(z)|\ge a} |h_n(z)|\,\mu_n(dz) &\le \Big(\frac{1}{a}\Big)^{\rho-1} \int_{\mathcal{M}} |h_n(z)|^\rho \,\mu_n(dz) \;=\; \Big(\frac{1}{a}\Big)^{\rho-1} \int_{\mathbb{R}^s} |h_n(z)|^\rho \,\mu_n(dz)\\ &\le \Big(\frac{1}{a}\Big)^{\rho-1} \int_{\mathbb{R}^s} \big(\kappa h(z) + \kappa h(-Ax_n)\big)^\rho \;\mu_n(dz)\\ &\le \Big(\frac{1}{a}\Big)^{\rho-1} \int_{\mathbb{R}^s} \big(\kappa h(z) + \kappa_o\big)^\rho \;\mu_n(dz),\end{aligned}$$

where we have used (A3.1) and (A3.3) in the second inequality, and where in the third inequality $\kappa_o > 0$ is a suitable constant selected according to (A3.4). The estimate continues as follows:

$$\begin{aligned}&\le \Big(\frac{1}{a}\Big)^{\rho-1} \Big((2\kappa_o)^\rho \;+\; (2\kappa)^\rho \int_{\mathbb{R}^s} h(z)^\rho \,\mu_n(dz)\Big)\\ &\le \Big(\frac{1}{a}\Big)^{\rho-1} \big((2\kappa_o)^\rho \;+\; (2\kappa)^\rho \bar{h}\big).\end{aligned}$$

This establishes the desired uniform integrability (2.2). The proof is completed by using (2.2) and Theorem 5.4 of [3]; see also the proof of Proposition 3.6 (ii) in [18]. □

3. The bounded convex case. The structural analysis of $P(\mu, \lambda)$ can be pushed ahead under suitable boundedness and convexity assumptions. We assume the following:

(A4) The set $Y_\lambda := \{y \in \mathbb{R}^m_+ \,:\, \lambda(H_j(y)) \geq p_j, j = 1, \ldots, d\}$ is convex for any $\lambda \in \Lambda$, and $Y := \cup_{\lambda\in\Lambda} Y_\lambda$ is bounded.

Sufficient conditions for the convexity in (A4) can be formulated for the case that the mappings H_j all have closed convex graphs—see (A.5) below, with the help of r-concave and logarithmic concave probability measures. The uniform distribution, the nondegenerate multivariate normal distribution, the multidimensional Dirichlet distribution, and the multivariate student and Pareto distributions all fit into this framework; for details consult [8] and [11].

LEMMA 8. *Under* (A4) *the following hold for any* $\lambda \in \Lambda$*:*

(i) $Q(t, \lambda) \in \mathbb{R}$*, whenever* $t \in W(Y_\lambda)$.

(ii) $Q(\cdot, \lambda)$ *is continuous on* $\operatorname{int} W(Y_\lambda)$.

(iii) *There exists a constant* $\bar\kappa > 0$*, uniformly for all* $\lambda \in \Lambda$*, such that* $|Q(t, \lambda)| \leq \bar\kappa$ *for all* $t \in W(Y_\lambda)$.

Proof. Assertions (i) and (iii) immediately follow from the compactness of each Y_λ, the boundedness of Y, and the continuity of $q^T y$. To confirm (ii), note that $Q(\cdot, \lambda)$ has to be convex as the value function of a convex program with parameters in the right-hand side. As a (real-valued) convex function (in finite dimension) $Q(\cdot, \lambda)$ is thus continuous on the open set $\operatorname{int} W(Y_\lambda)$. □

Remark 9. For fixed $\lambda \in \Lambda$ the above lemma immediately gives rise to specifications of Lemma 2, Lemma 3, and Proposition 7. Indeed, (A1) holds with $\mathcal{R}_\lambda = W(Y_\lambda)$ and $\mathcal{N}_\lambda \subseteq \partial W(Y_\lambda)$, where ∂ denotes the boundary. Since $W(Y_\lambda)$ is convex, its boundary has Lebesgue measure zero; see, e.g., [9]. Assumption (A3) can obviously be fulfilled with $h(t) \equiv \bar\kappa$.

We turn our attention to analyzing $P(\mu, \lambda)$ with both μ and λ variable. To this end we define the following distance (discrepancy) on Λ:

$$\alpha(\mu, \nu) \;:=\; \max_{j=1,\ldots,d} \sup\{|\mu(B) - \nu(B)| \,:\, B \in \mathcal{B}_j\},$$

where each $\mathcal{B}_j, j = 1, \ldots, d$, is a class of Borel sets of $\mathbb{R}^s$ such that $\{H_j(y) \,:\, y \in \mathbb{R}^m\} \subseteq \mathcal{B}_j$ and that α forms a metric. We fix a vector $\bar p$ of probability levels and introduce the notation

$$\begin{aligned}\Psi(t, \lambda) &:= \operatorname{argmin}\{q^T y \,:\, Wy = t,\, y \geq 0,\, \lambda(H_j(y)) \geq \bar p_j,\, j = 1, \ldots, d\},\\ C_p(t, \lambda) &:= \{y \in \mathbb{R}^m \,:\, Wy = t,\, y \geq 0,\, \lambda(H_j(y)) \geq p_j,\, j = 1, \ldots, d\}.\end{aligned}$$

Now the following holds.

LEMMA 10. *Assume* (A4)*, and fix some* $(t_o, \lambda_o) \in \mathbb{R}^s \times \Lambda$ *such that* $\Psi(t_o, \lambda_o) \neq \emptyset$*. For each* $y_o \in \Psi(t_o, \lambda_o)$ *let the function*

$$(\xi, t, p) \;\mapsto\; \inf\{\|\xi - y\| \,:\, y \in C_p(t, \lambda_o)\} \tag{3.1}$$

be Lipschitz continuous on some neighborhood of $(y_o, t_o, \bar p)$.

Then the multifunction Ψ *(acting from the metric space* $[(\mathbb{R}^s, \Lambda), (\|.\|, \alpha)]$ *into the subsets of* $\mathbb{R}^m$*) is upper semicontinuous at* (t_o, λ_o)*. Furthermore, there exist* $L > 0$

and $\delta > 0$ such that

$$\Psi(t,\lambda) \neq \emptyset \qquad \textit{and}$$
$$|Q(t,\lambda) - Q(t_o,\lambda_o)| \;\leq\; L \cdot (\|t - t_o\| + \alpha(\lambda,\lambda_o))$$

whenever $\|t - t_o\| + \alpha(\lambda,\lambda_o) < \delta$.

Proof. The proof employs Lemma A.2 from the appendix of [15]. It is sufficient to verify assumption (c) of that lemma. In our setting, this condition reads as follows: For each $y_o \in \Psi(t_o,\lambda_o)$ there exist a neighborhood $U = U(y_o)$ and positive reals δ_o and l such that the following hold for all $(t,\lambda) \in \mathbb{R}^s \times \Lambda$ with $\|t - t_o\| + \alpha(\lambda,\lambda_o) < \delta_o$:

$$y \in C_{\bar p}(t_o,\lambda_o) \cap U \quad \text{implies} \quad d(y, C_{\bar p}(t,\lambda)) \;\leq\; l \cdot (\|t - t_o\| + \alpha(\lambda,\lambda_o)) \tag{3.2}$$

and

$$y \in C_{\bar p}(t,\lambda) \cap U \quad \text{implies} \quad d(y, C_{\bar p}(t_o,\lambda_o)) \;\leq\; l \cdot (\|t - t_o\| + \alpha(\lambda,\lambda_o)). \tag{3.3}$$

Let $y_o \in \Psi(t_o,\lambda_o)$ be fixed. Assumption (3.1) is equivalent to the existence of neighborhoods U of y_o and V of $(t_o,\bar p)$ and of a constant $l > 0$ such that $d(y, C_p(t,\lambda_o))$ is finite and

$$d(y, C_{p^1}(t_1,\lambda_o)) \;\leq\; d(y, C_{p^2}(t_2,\lambda_o)) \;+\; l \cdot (\|t_1 - t_2\| + \|p^1 - p^2\|) \tag{3.4}$$

for all $y \in U$ and all $(t,p), (t_1,p^1), (t_2,p^2) \in V$.

Let $\delta_o > 0$ such that

$$\{(t,p) \;:\; \|t - t_o\| + \|p - \bar p\| \leq \delta_o\} \subseteq V$$

and let $\lambda \in \Lambda$ such that $\alpha(\lambda,\lambda_o) \leq \delta_o$. Then by definition of the distance α the following inclusions hold:

$$C_{\bar p + 1\alpha(\lambda,\lambda_o)}(t,\lambda_o) \;\subseteq\; C_{\bar p}(t,\lambda) \;\subseteq\; C_{\bar p - 1\alpha(\lambda,\lambda_o)}(t,\lambda_o). \tag{3.5}$$

(Here, 1 denotes the vector of all ones.) Let $y \in U$, and set in (3.4)

$$p^1 := \bar p + 1\alpha(\lambda,\lambda_o), \quad p^2 := \bar p, \quad t_1 := t, \quad \text{and} \quad t_2 := t_o.$$

Then (3.5) and (3.4) imply[1]

$$\begin{aligned} d(y, C_{\bar p}(t,\lambda)) &\leq d(y, C_{\bar p + 1\alpha(\lambda,\lambda_o)}(t,\lambda_o)) \\ &\leq d(y, C_{\bar p}(t_o,\lambda_o)) + l \cdot (\|t - t_o\| + \alpha(\lambda,\lambda_o)), \end{aligned}$$

yielding (3.2). Now set in (3.4)

$$p^1 := \bar p \;\;, \quad p^2 := \bar p - 1\alpha(\lambda,\lambda_o), \quad t_1 := t_o, \quad \text{and} \quad t_2 := t.$$

It follows that

$$d(y, C_{\bar p}(t_o,\lambda_o)) \;\leq\; d(y, C_{\bar p - 1\alpha(\lambda,\lambda_o)}(t,\lambda_o)) + l \cdot (\|t - t_o\| + \alpha(\lambda,\lambda_o)).$$

[1] To avoid further multiplicative constants we assume that the norm in $\|p^1 - p^2\|$ is $\|.\|_\infty$.

Together with (3.5) we obtain

$$d(y, C_{\bar{p}}(t_o, \lambda_o)) \;\leq\; d(y, C_{\bar{p}}(t, \lambda)) + l \cdot (\|t - t_o\| + \alpha(\lambda, \lambda_o)),$$

yielding (3.3) and completing the proof. □

Remark 11. The above result extends Theorem 3.2 in [15] to the present situation. As in [15] the result can be further extended to the nonconvex case, provided a suitable notion of local optimality (complete local minimizing set) is employed.

Remark 12. The essential assumption in Lemma 10 is (3.1). It can be equivalently restated as follows: The multifunction $(t, p) \mapsto C_p(t, \lambda_o)$ is pseudo-Lipschitzian in the sense of [13] at each $(y_o, t_o, \bar{p}) \in \Psi(t_o, \lambda_o) \times \{(t_o, \bar{p})\}$. Sufficient conditions for multifunctions to be pseudo-Lipschitzian have been derived in the literature. The following is a popular one, often called the Robinson–Ursescu theorem (here already adapted to our setting): Let Γ be a multifunction from $\mathbb{R}^s \times \mathbb{R}^d$ into $\mathbb{R}^m$ with closed convex graph. Then Γ is pseudo-Lipschitzian at each pair $(y_o, t_o, \bar{p}) \in \Gamma(t_o, \bar{p}) \times \mathbb{R}^{s+d}$ with $(t_o, \bar{p}) \in \operatorname{int}(\operatorname{dom}\Gamma)$, where $\operatorname{dom}\Gamma \;=\; \{(t, p) \in \mathbb{R}^{s+d} \,:\, \Gamma(t, p) \neq \emptyset\}$.

The above remark gives rise to the following assumption.

(A5) Each $H_j, j = 1, \ldots, d$, has a closed convex graph.

LEMMA 13. *Assume* (A4), (A5) *and let* $\lambda_o \in \Lambda$. *If* $t_o \in \operatorname{int} W(Y_{\lambda_o})$, *then*

$$(t_o, \bar{p}) \in \operatorname{int}(\operatorname{dom} C_{(\cdot)}(\cdot, \lambda_o)).$$

Proof. If $t_o \in \operatorname{int} W(Y_{\lambda_o})$, then there exists an open ball $B(t_o, r)$ such that $B(t_o, r) \subseteq \operatorname{int} W(Y_{\lambda_o})$. By continuity, the preimage under W of that ball is open, and the preimage is obviously contained in Y_{λ_o}. Let $y_s \in W^{-1}(\{t_o\})$. Then $y_s \in W^{-1}(B(t_o, r)) \subseteq \operatorname{int} Y_{\lambda_o}$. By convexity, this implies that $y_s > 0$ and $\lambda_o(H_j(y_s)) > \bar{p}_j$ for all j. Hence $(t_o, \bar{p}) \in \operatorname{int}(\operatorname{dom} C_{(\cdot)}(\cdot, \lambda_o))$. □

The combination of Lemmas 10 and 13 produces the following corollary.

COROLLARY 14. *Assume* (A4), (A5) *and let* $\lambda_o \in \Lambda$ *and* $t_o \in \operatorname{int} W(Y_{\lambda_o})$. *Then there exist* $L > 0$ *and* $\delta > 0$ *such that*

$$|Q(t, \lambda) - Q(t_o, \lambda_o)| \;\leq\; L \cdot (\|t - t_o\| + \alpha(\lambda, \lambda_o)) \quad \textit{whenever } \|t - t_o\| + \alpha(\lambda, \lambda_o) < \delta.$$

We are now in a position to formulate a sufficient condition for joint continuity of G. Assumption (A2*) adapts as follows to the setting of the present section:

(A2*) there exists an open set $C^* \supseteq C$ such that $\displaystyle\bigcup_{\mu \in \Delta} \operatorname{supp} \mu \subseteq \bigcap_{\lambda \in \Lambda} \bigcap_{x \in C^*} \{Ax + W(Y_\lambda)\}$.

PROPOSITION 15. *Assume* (A2*), (A4), (A5) *and let* A *have full row rank. Then* $G : C^* \times \Delta \times \Lambda \longrightarrow \mathbb{R}$ *is continuous at any triplet* $(x, \mu, \lambda) \in C \times \Delta \times \Lambda$ *fulfilling* $\mu\big(Ax + \partial W(Y_\lambda)\big) = 0$.

Proof. Let (x, μ, λ) be as in the assumption, $x_n \longrightarrow x$, $\mu_n \xrightarrow{w} \mu$, $\lambda_n \xrightarrow{\alpha} \lambda$, and assume without loss of generality that $x_n \in C^*$ for all n. In analogy to the proof of Proposition 7 let $\mathcal{M} := \cup_{\nu \in \Delta} \operatorname{supp} \nu$, define functions $h_n, h_o : \mathcal{M} \to \mathbb{R}$ by $h_n(z) := Q(z - Ax_n, \lambda_n)$ and $h_o(z) := Q(z - Ax, \lambda)$, and consider the set

$$E(x) \;:=\; \{z \in \mathcal{M} \,:\, \exists z_n \to z \text{ with } h_n(z_n) \not\to h_o(z)\}.$$

Adapting the corresponding argument from the proof of Proposition 7 and taking into account Corollary 14, we obtain that $E(x) \subseteq Ax + \partial W(Y_\lambda)$. The assumption

on (x, μ, λ) thus implies $\mu(E(x)) = 0$, and Rubin's theorem on weak convergence of image measures yields

$$\mu_n \circ h_n^{-1} \xrightarrow{w} \mu \circ h_o^{-1}. \tag{3.6}$$

With $\bar{\kappa}$ according to Lemma 8(iii) we consider the bounded continuous function

$$g(\tau) := \begin{cases} \bar{\kappa}, & \tau \geq \bar{\kappa}, \\ \tau, & -\bar{\kappa} \leq \tau \leq \bar{\kappa}, \\ -\bar{\kappa}, & \tau \leq -\bar{\kappa}. \end{cases}$$

By (3.6) it holds that

$$\int_{\mathbb{R}} g(\tau)\, \mu_n \circ h_n^{-1}(d\tau) \longrightarrow \int_{\mathbb{R}} g(\tau)\, \mu \circ h_o^{-1}(d\tau).$$

Changing variables yields

$$\int_{\mathcal{M}} h_n(z)\, \mu_n(dz) \longrightarrow \int_{\mathcal{M}} h_o(z)\, \mu(dz),$$

proving the assertion. □

Since $\partial W(Y_\lambda)$ always has Lebesgue measure zero, we obtain the following corollary; see also Remarks 4 and 9.

COROLLARY 16. *Assume* (A2*), (A4), (A5). *Let A have full row rank and let μ have a density. Then $G : C^* \times \Delta \times \Lambda \longrightarrow \mathbb{R}$ is continuous at any $(x, \mu, \lambda) \in C \times \{\mu\} \times \Lambda$.*

The joint continuity established in Proposition 15 and Corollary 16 paves the way for qualitative stability of $P(\mu, \lambda)$ under perturbations of (μ, λ). To illustrate this, we conclude the paper with a corresponding result based on Corollary 16. To this end, we write $P(\mu, \lambda)$ in a more compact form as

$$P(\mu, \lambda) \qquad \min\{c^T x + G(x, \mu, \lambda) \ : \ x \in C\}, \quad (\mu, \lambda) \in \Delta \times \Lambda.$$

Notice that in the setting of Corollary 16, $G(\cdot, \mu', \lambda')$ is convex for all $(\mu', \lambda') \in \Delta \times \Lambda$. We assume that, in addition, C is convex, and we introduce the following notation:

$$\begin{aligned} \phi(\mu, \lambda) &:= \inf\{c^T x + G(x, \mu, \lambda) \ : \ x \in C\}, \\ \psi(\mu, \lambda) &:= \{x \in C \ : \ c^T x + G(x, \mu, \lambda) = \phi(\mu, \lambda)\}. \end{aligned}$$

PROPOSITION 17. *Assume* (A2*), (A4), (A5). *Let A have full row rank, let $(\mu, \lambda) \in \Delta \times \Lambda$, and let μ have a density. Suppose further that $\psi(\mu, \lambda)$ is nonempty and bounded. Then it holds that*

(i) *the function $\phi : \Delta \times \Lambda \longrightarrow \mathbb{R}$ is continuous at (μ, λ);*

(ii) *the multifunction $\psi : \Delta \times \Lambda \longrightarrow 2^{\mathbb{R}^m}$ is upper semicontinuous at (μ, λ).*

With joint continuity of G in (x, μ, λ) established, the proof of the above proposition follows the lines of Berge's theory as, for instance, in the proof of Proposition 4.2.2 in [2].

REFERENCES

[1] Z. ARTSTEIN AND R. J.-B. WETS, *Stability results for stochastic programs and sensors, allowing for discontinuous objective functions*, SIAM J. Optim., 4 (1994), pp. 537–550.

[2] B. Bank, J. Guddat, D. Klatte, B. Kummer, and K. Tammer, *Nonlinear Parametric Optimization*, Akademie-Verlag, Berlin, 1982.
[3] P. Billingsley, *Convergence of Probability Measures*, Wiley, New York, 1968.
[4] P. Billingsley, *Probability and Measure*, 2nd ed., Wiley, New York, 1986.
[5] J. Dupačová, *Stability and sensitivity analysis for stochastic programming*, Ann. Oper. Res., 27 (1990), pp. 115–142.
[6] R. Henrion, and W. Römisch, *Hölder and Lipschitz stability of solution sets in programs with probabilistic constraints*, Math. Program., 100 (2004), pp. 589–611.
[7] P. Kall, *On approximations and stability in stochastic programming*, in Parametric Optimization and Related Topics, J. Guddat, H. Th. Jongen, B. Kummer, and F. Nožička, eds., Akademie-Verlag, Berlin, 1987, pp. 387–407.
[8] A. I. Kibzun and Y. S. Kan, *Stochastic Programming Problems with Probability and Quantile Functions*, Wiley-Interscience Series in Systems and Optimization, Wiley, New York, 1996.
[9] R. Lang, *A note on the measurability of convex sets*, Arch. Math. (Basel), 47 (1986), pp. 90–92.
[10] F. V. Louveaux and R. Schultz, *Stochastic integer programming*, in Stochastic Programming, Handbooks Oper. Res. Management Sci. 10, A. Ruszczyński and A. Shapiro, eds., Elsevier, Amsterdam, 2003, pp. 213–266.
[11] A. Prékopa, *Stochastic Programming*, Kluwer Academic, Dordrecht, The Netherlands, 1995.
[12] S. M. Robinson and R. J.-B. Wets, *Stability in two-stage stochastic programming*, SIAM J. Control Optim., 25 (1987), pp. 1409–1416.
[13] R. T. Rockafellar, *Lipschitzian properties of multifunctions*, Nonlinear Anal., 9 (1985), pp. 867–885.
[14] W. Römisch, *Stability of stochastic programming problems*, in Stochastic Programming, Handbooks Oper. Res. Management Sci. 10, A. Ruszczyński and A. Shapiro, eds., Elsevier, Amsterdam, 2003, pp. 483–554.
[15] W. Römisch and R. Schultz, *Stability analysis for stochastic programs*, Ann. Oper. Res., 30 (1991), pp. 241–266.
[16] R. Rubisch, *Ein zweistufiges stochastisches Optimierungsmodell aus der Kraftwerkseinsatzplanung mit Wahrscheinlichkeitsrestriktionen in der zweiten Stufe*, Master Thesis, Preprint 637, Department of Mathematics, University of Duisburg–Essen, Duisburg, Germany, 2006; http://www.uni-due.de/mathematik/d_preprints06.shtml.
[17] A. Ruszczyński and A. Shapiro, eds., *Stochastic Programming*, Handbooks Oper. Res. Management Sci. 10, Elsevier, Amsterdam, 2003.
[18] R. Schultz, *Mixed-integer value functions in stochastic programming*, in Combinatorial Optimization: Eureka, You Shrink!, Papers Dedicated to Jack Edmonds, M. Jünger, G. Reinelt, and G. Rinaldi, eds., Springer-Verlag, Berlin, 2003, pp. 171–184.
[19] R. Schultz, *Some aspects of stability in stochastic programming*, Ann. Oper. Res., 100 (2000), pp. 55–84.
[20] A. Shapiro, *Quantitative stability in stochastic programming*, Math. Programming, 67 (1994), pp. 99–108.
[21] A. Shapiro and T. Homem-de-Mello, *On the rate of convergence of optimal solutions of Monte Carlo approximations of stochastic programs*, SIAM J. Optim., 11 (2000), pp. 70–86.
[22] R. J.-B. Wets, *Stochastic programming: Solution techniques and approximation schemes*, in Mathematical Programming: State of the Art, Bonn 1982, A. Bachem, M. Grötschel, and B. Korte, eds., Springer-Verlag, Berlin, 1983, pp. 560–603.
[23] R. J.-B. Wets, *Stochastic Programming*, in Optimization, Handbooks Oper. Res. Management Sci. 1, G. L. Nemhauser, A. H. G. Rinnooy Kan, and M. J. Todd, eds., North–Holland, Amsterdam, 1989, pp. 573–629.
[24] R. J.-B. Wets, *Challenges in stochastic programming*, Math. Programming, 75 (1996), pp. 115–135.

SIAM J. OPTIM.
Vol. 18, No. 3, pp. 789–809

FITZPATRICK FUNCTIONS AND CONTINUOUS LINEAR MONOTONE OPERATORS*

HEINZ H. BAUSCHKE†, JONATHAN M. BORWEIN‡, AND XIANFU WANG†

Abstract. The notion of a maximal monotone operator is crucial in optimization as it captures both the subdifferential operator of a convex, lower semicontinuous, and proper function and any (not necessarily symmetric) continuous linear positive operator. It was recently discovered that most fundamental results on maximal monotone operators allow simpler proofs utilizing Fitzpatrick functions. In this paper, we study Fitzpatrick functions of continuous linear monotone operators defined on a Hilbert space. A novel characterization of skew operators is presented. A result by Brézis and Haraux is reproved using the Fitzpatrick function. We investigate the Fitzpatrick function of the sum of two operators, and we show that a known upper bound is actually exact in finite-dimensional and more general settings. Cyclic monotonicity properties are also analyzed, and closed forms of the Fitzpatrick functions of all orders are provided for all rotators in the Euclidean plane.

Key words. cyclic monotonicity, Fitzpatrick family, Fitzpatrick function, linear operator, maximal monotone operator, Moore–Penrose inverse, paramonotone operator, rotator

AMS subject classifications. 47H05, 47B25, 47B65, 90C25

DOI. 10.1137/060655468

1. Introduction.

Throughout this paper, we assume that

(1) X is a real Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and induced norm $\| \cdot \|$.

Recall that a set-valued operator $A\colon X \to 2^X$ is *monotone* if

$$\left.\begin{array}{l}(x,u) \in \operatorname{gra} A\\ (y,v) \in \operatorname{gra} A\end{array}\right\} \quad \Rightarrow \quad \langle x-y, u-v\rangle \geq 0, \tag{2}$$

where $\operatorname{gra} A = \big\{(x,u) \in X \times X \mid u \in Ax\big\}$ denotes the *graph* of A. The notion of a monotone operator is central to modern optimization and analysis [9, 10, 33, 34, 35, 36, 43]. Of particular importance are *maximal monotone operators*, i.e., monotone operators with graphs that cannot be enlarged without destroying monotonicity. Recently, several fundamental results on monotone operators have found—sometimes dramatically simpler—new proofs by utilizing *Fitzpatrick functions* [8, 9, 29, 39, 41, 42]. The Fitzpatrick function was first introduced by Fitzpatrick to study monotone operators via convex analysis [17]; see also [2, 6, 12, 13, 14, 15, 18, 24, 25, 31, 37, 38, 40]. The key classes of maximal monotone operators are subdifferential operators of proper, lower semicontinuous, and convex functions [32] and continuous, linear, and monotone operators. The former class is very well understood [36]; the latter class is the topic of this paper.

*Received by the editors March 28, 2006; accepted for publication (in revised form) November 21, 2006; published electronically October 4, 2007. This work was partially supported by the Natural Sciences and Engineering Research Council of Canada.

http://www.siam.org/journals/siopt/18-3/65546.html

†Department of Mathematics, Irving K. Barber School, University of British Columbia Okanagan, Kelowna, BC V1V 1V7, Canada (heinz.bauschke@ubc.ca, shawn.wang@ubc.ca). The research of the first author was partially supported by the Canada Research Chair Program.

‡Faculty of Computer Science, Dalhousie University, 6050 University Avenue, Halifax, NS B3H 1W5, Canada (jborwein@cs.dal.ca). The research of this author was partially supported by the Canada Research Chair Program.

It is well known that continuous, linear, and monotone operators are automatically maximal monotone (see, e.g., [36, p. 30]); see also [1, 5, 7, 30, 36] for additional works on monotone-operator-theoretic properties of linear operators. Let $A\colon X \to X$ be continuous and linear. Linearity and (2) yield

$$A \text{ is monotone} \quad \Leftrightarrow \quad (\forall x \in X)\ \langle x, Ax\rangle \geq 0. \tag{3}$$

Thus monotonicity is determined solely by the behavior of the *symmetric part* of A. We now recall the relevant notions.

DEFINITION 1.1 (symmetric and skew part). *Let $A\colon X \to X$ be continuous and linear. Then $A_+ = \frac{1}{2}A + \frac{1}{2}A^*$ is the* symmetric part *of A, and $A_\circ = A - A_+ = \frac{1}{2}A - \frac{1}{2}A^*$ is the* skew part *of A.*

The next result is clear.

PROPOSITION 1.2. *Let $A\colon X \to X$ be continuous and linear. Then A is monotone if and only if A_+ is monotone.*

Let us define the Fitzpatrick function [17] for linear operators.

DEFINITION 1.3 (Fitzpatrick function). *Let $A\colon X \to X$ be continuous and linear. The* Fitzpatrick function *of A is*

$$F_A\colon X \times X \to \left]-\infty, +\infty\right]\colon (x,u) \mapsto \sup_{y\in X}\big(\langle x, Ay\rangle + \langle y, u\rangle - \langle y, Ay\rangle\big). \tag{4}$$

Before we survey some fundamental results concerning the Fitzpatrick function of a linear operator, we need to briefly explain our notation. We shall utilize throughout this paper notation and results that are standard in convex analysis and monotone operator theory. See [9, 33, 35, 36, 43] for comprehensive references. The *Fenchel conjugate* and *domain* of a function f is denoted by f^* and $\operatorname{dom} f$, respectively. The *ball* of radius ρ centered at x is denoted by $B(x;\rho)$. The *closure*, the *interior*, and the *indicator function* of a set $S \subseteq X$ are written as $\overline{S}$, $\operatorname{int} S$, and ι_S, respectively. For a continuous and linear operator $A\colon X \to X$, the *kernel* (also known as null space) of A is denoted by $\ker A$ and the *range* by $\operatorname{ran} A$. The identity operator is written as Id. If A is monotone and symmetric, it will occasionally be convenient to use the notation

$$(\forall x \in X)(\forall y \in X) \quad \langle x, y\rangle_A = \langle x, Ay\rangle \ \text{ and } \ \|x\|_A = \sqrt{\langle x, x\rangle_A} = \|\sqrt{A}x\|, \tag{5}$$

where $\sqrt{A}$ denotes the *square root* of A [23, section 9.4].

FACT 1.4 (see [17]). *Let $A\colon X \to X$ be continuous, linear, and monotone. Then*

(i) *F_A is convex, lower semicontinuous, and proper;*

(ii) *$F_A = \langle\cdot,\cdot\rangle$ on $\operatorname{gra} A$, and $F_A > \langle\cdot,\cdot\rangle$ outside $\operatorname{gra} A$;*

(iii) $\big(\forall (x,u) \in X \times X\big)\ F_A(x,u) \leq F_A^*(u,x) = \big(\iota_{\operatorname{gra} A} + \langle\cdot,\cdot\rangle\big)^{**}(x,u)$.

Fact 1.4(ii) motivates the following definition (see also [14]).

DEFINITION 1.5 (Fitzpatrick family). *Let $A\colon X \to X$ be continuous, linear, and monotone. The* Fitzpatrick family $\mathcal{F}_A$ *consists of all functions $F\colon X \times X \to \left]-\infty, +\infty\right]$ such that F is convex, lower semicontinuous, $F \geq \langle\cdot,\cdot\rangle$, and $F = \langle\cdot,\cdot\rangle$ on $\operatorname{gra} A$.*

FACT 1.6 (see [17]). *Let $A\colon X \to X$ be continuous, linear, and monotone. Then for every $(x,u) \in X \times X$,*

$$F_A(x,u) = \min_{F\in\mathcal{F}_A} F(x,u) \quad \text{and} \quad F_A^*(u,x) = \max_{F\in\mathcal{F}_A} F(x,u). \tag{6}$$

The plan for the remainder of the paper is as follows.

- In section 2, we describe completely the Fitzpatrick function and its conjugate (Theorem 2.3). Some examples and a new *characterization of skew operators* in terms of the Fitzpatrick family (Theorem 2.9) are provided.
- The *range* of a continuous, linear, and monotone operator is studied in section 3 and compared to the range of the adjoint. The closures of these two ranges coincide; however, the Volterra integral operator (Example 3.3) illustrates that the ranges themselves can differ.
- Section 4 deals with *rectangular*—also known as *property* $(*)$ *monotone*—operators, a class of operators introduced by Brézis and Haraux [11]. We state their main result and discuss some useful consequences. We also provide a characterization of rectangular operators in terms of their symmetric and skew parts (Corollary 4.10). This allows us to make a connection with *paramonotone operators* (Remark 4.11). A result by Brézis and Haraux is re-proved using the Fitzpatrick function (Theorem 4.12).
- We turn to the *Fitzpatrick function of the sum* in section 5. No general formula is known; in fact, Fitzpatrick posed this as an open problem (see [17, Problem 5.4]). We present a partial solution to his problem by showing that a known upper bound is actually exact in finite-dimensional spaces (Corollary 5.7) as well as in more general settings (Theorems 5.3 and 5.4 and Corollary 5.6).
- *Cyclic monotonicity* is a quantitative refinement of monotonicity that can be captured with higher-order Fitzpatrick functions. We begin in section 6 by reviewing known results about these functions. We then present a new closed form (Example 6.4), a novel recursion formula (Theorem 6.5), and a localization of the domain (Corollary 6.7).
- Finally, in section 7, we study cyclic monotonicity properties and higher-order Fitzpatrick functions of *rotators* in the Euclidean plane. Complete characterizations of n-cyclic monotonicity and explicit formulas for the Fitzpatrick functions are provided in all possible cases (Theorem 7.8). This considerably extends previously known results [2, section 4].

2. The Fitzpatrick function and skew operators. The Fitzpatrick function of a continuous linear operator will be formulated in terms of a quadratic function that we present next.

DEFINITION 2.1 (quadratic function). *Let $A\colon X \to X$ be continuous, linear, and symmetric. Then we set $q_A\colon X \to \mathbb{R}\colon x \mapsto \frac{1}{2}\langle x, Ax\rangle$.*

FACT 2.2. *Let $A\colon X \to X$ be continuous, linear, and symmetric. Then*

$$q_A \text{ is convex} \Leftrightarrow A \text{ is monotone.} \tag{7}$$

In this case, the following is true:

(i) $\nabla q_A = A$.
(ii) $q_A^* \circ A = q_A$.
(iii) $\operatorname{ran} A \subseteq \operatorname{dom} q_A^* \subseteq \overline{\operatorname{ran}}\, A$.
(iv) $q_A^* \geq 0$ *and* $(\forall u \in X)(\forall \rho \in \mathbb{R} \smallsetminus \{0\})\ q_A^*(\rho u) = \rho^2 q_A^*(u)$. *Consequently,* $\operatorname{dom} q_A^*$ *is a subspace.*
(v) *If* $\operatorname{ran} A$ *is closed, then* $q_A^* = \iota_{\operatorname{ran} A} + q_{A^\dagger}$, *where* $A^\dagger$ *is the* Moore–Penrose inverse [20] *of* A.
(vi) *If* A *is bijective, then* $q_A^* = q_{A^{-1}}$.

Proof. (See also [3, Proposition 12.3.6].) (i) and (ii) See [5, Theorem 3.6.(i)]. See also [30, Theorem 5.1] for a considerably more general version of (i). (iii) See [4, Fact

2.2(iii)]. (iv) The proof of item (iv) is elementary. (v) See [6, Proposition 3.7(iv)]. (vi) is clear from (v). ∎

THEOREM 2.3. *Let $A\colon X \to X$ be continuous, linear, and monotone. Then*

(i) $\big(\forall (x,u) \in X \times X\big)\quad F_A(x,u) = 2q^*_{A_+}\big(\frac{1}{2}u + \frac{1}{2}A^*x\big) = F_{A_+}(x, u - A_\circ x)$;

(ii) $\operatorname{ran} A_+ \subseteq (A^* \oplus \operatorname{Id})(\operatorname{dom} F_A) = \operatorname{dom} q^*_{A_+} \subseteq \overline{\operatorname{ran}}\, A_+$;

(iii) $\big(\forall (u,x) \in X \times X\big)\quad F_A^*(u,x) = \iota_{\operatorname{gra} A}(x,u) + \langle x, Ax\rangle$.

Proof. Fix $(x,u) \in X \times X$. (i) This follows from

$$\begin{aligned} F_A(x,u) &= \sup_{y\in X} \langle x, Ay\rangle + \langle y, u\rangle - \langle y, Ay\rangle = 2\sup_{y\in X}\langle y, \tfrac{1}{2}u + \tfrac{1}{2}A^*x\rangle - q_{A_+}(y) \\ &= 2q^*_{A_+}\big(\tfrac{1}{2}u + \tfrac{1}{2}A^*x\big) = 2q^*_{A_+}\big(\tfrac{1}{2}(u - A_\circ x) + \tfrac{1}{2}A_+x\big) = F_{A_+}(x, u - A_\circ x). \end{aligned} \tag{8}$$

(ii) The equality is a consequence of (i), and the inclusions are then clear from Fact 2.2(iii). (iii) This follows from Fact 1.4(iii) and the fact that the function $(u,x) \mapsto \iota_{\operatorname{gra} A}(x,u) + \langle x, u\rangle = \iota_{\operatorname{gra} A}(x,u) + \langle x, Ax\rangle$ is already convex, lower semicontinuous, and proper. ∎

The next two results play a role in the proof of Theorem 5.4 below.

Example 2.4. Let $A\colon X \to X$ be continuous, linear, monotone, and symmetric. Then

$$\big(\forall x \in X\big)\big(\forall y \in X\big)\quad F_A(x,Ay) = \tfrac{1}{4}\langle x+y, A(x+y)\rangle. \tag{9}$$

Proof. Take $x \in X$ and $y \in X$. Using Theorem 2.3(i) and Fact 2.2(ii), we obtain

$$F_A(x,Ay) = 2q^*_A\big(\tfrac{1}{2}Ay + \tfrac{1}{2}Ax\big) = 2q_A\big(\tfrac{1}{2}x + \tfrac{1}{2}y\big) = \tfrac{1}{4}\langle x+y, A(x+y)\rangle, \tag{10}$$

as required. ∎

Example 2.5 (closed range symmetric operator). Let $A\colon X \to X$ be continuous, linear, monotone, and symmetric such that $\operatorname{ran} A$ is closed. Then

$$\begin{aligned} \big(\forall (x,u) \in X \times X\big)\quad F_A(x,u) &= \iota_{\operatorname{ran} A}(u) + \tfrac{1}{4}\big(\langle x, Ax\rangle + 2\langle x, u\rangle + \langle A^\dagger u, u\rangle\big) \\ &= \iota_{\operatorname{ran} A}(u) + \tfrac{1}{4}\|x + A^\dagger u\|_A^2 \end{aligned} \tag{11}$$

and hence

$$\operatorname{dom} F_A = X \times \operatorname{ran} A. \tag{12}$$

Proof. Fix $(x,u) \in X \times X$. Using Theorem 2.3(i), Fact 2.2(v), and standard properties of the Moore–Penrose inverse [20], we deduce that

$$\begin{aligned} F_A(x,u) &= 2q^*_A\big(\tfrac{1}{2}u + \tfrac{1}{2}Ax\big) \\ &= 2\iota_{\operatorname{ran} A}\big(\tfrac{1}{2}u + \tfrac{1}{2}Ax\big) + 2q_{A^\dagger}\big(\tfrac{1}{2}u + \tfrac{1}{2}Ax\big) \\ &= \iota_{\operatorname{ran} A}(u) + \langle A^\dagger(\tfrac{1}{2}u + \tfrac{1}{2}Ax), \tfrac{1}{2}u + \tfrac{1}{2}Ax\rangle \\ &= \iota_{\operatorname{ran} A}(u) + \tfrac{1}{4}\big(\langle A^\dagger u, u\rangle + \langle A^\dagger u, Ax\rangle + \langle A^\dagger Ax, u\rangle + \langle A^\dagger Ax, Ax\rangle\big) \\ &= \iota_{\operatorname{ran} A}(u) + \tfrac{1}{4}\big(\langle A^\dagger u, u\rangle + \langle AA^\dagger u, x\rangle + \langle x, AA^\dagger u\rangle + \langle AA^\dagger Ax, x\rangle\big) \\ &= \iota_{\operatorname{ran} A}(u) + \tfrac{1}{4}\big(\langle x, Ax\rangle + 2\langle x, u\rangle + \langle A^\dagger u, u\rangle\big) \\ &= \iota_{\operatorname{ran} A}(u) + \tfrac{1}{4}\big(\langle x, Ax\rangle + \langle x, AA^\dagger u\rangle + \langle A^\dagger u, Ax\rangle + \langle A^\dagger u, AA^\dagger u\rangle\big) \\ &= \iota_{\operatorname{ran} A}(u) + \tfrac{1}{4}\langle x + A^\dagger u, A(x + A^\dagger u)\rangle \\ &= \iota_{\operatorname{ran} A}(u) + \tfrac{1}{4}\|x + A^\dagger u\|_A^2, \end{aligned} \tag{13}$$

as desired. ∎

Remark 2.6. Let A be as in Example 2.5. A referee pointed out that (12) can also be proved as follows: Take $(x,u)\in X\times X$. Utilizing Theorem 2.3(i) and Fact 2.2(iii), we have

$$(x,u)\in\operatorname{dom}F_A \;\Leftrightarrow\; \tfrac{1}{2}u+\tfrac{1}{2}Ax\in\operatorname{dom}q_A^*=\operatorname{ran}A \;\Leftrightarrow\; u\in\operatorname{ran}A. \tag{14}$$

Let us provide two further examples. The first one is related to [29, Example 1], while the second one generalizes [29, Example 3].

Example 2.7 (bijective symmetric operator). Let $A\colon X\to X$ be continuous, linear, monotone, symmetric, and bijective. Then

$$\begin{aligned}\big(\forall(x,u)\in X\times X\big)\quad F_A(x,u)&=\tfrac{1}{4}\big(\langle x,Ax\rangle+2\langle x,u\rangle+\langle A^{-1}u,u\rangle\big)\\&=\tfrac{1}{4}\|x+A^{-1}u\|_A^2.\end{aligned} \tag{15}$$

Proof. This is clear from Example 2.5. ∎

Example 2.8 (skew operator). Let $A\colon X\to X$ be continuous, linear, and skew. Then

$$\big(\forall(x,u)\in X\times X\big)\quad F_A(x,u)=F_A^*(u,x)=\iota_{\operatorname{gra}A}(x,u). \tag{16}$$

Proof. Since A is skew, $A^*=-A$, $A_+=0$ and thus $\operatorname{dom}q_{A_+}^*=\operatorname{ran}A_+=\{0\}$ is closed (Fact 2.2(iii)). Using Theorem 2.3(i), Fact 2.2(iv), and Theorem 2.3(iii), we obtain that

$$\begin{aligned}F_A(x,u)&=2q_{A_+}^*\big(\tfrac{1}{2}u+\tfrac{1}{2}A^*x\big)=2\iota_{\{0\}}\big(\tfrac{1}{2}u+\tfrac{1}{2}A^*x\big)\\&=2\iota_{\{0\}}\big(\tfrac{1}{2}u-\tfrac{1}{2}Ax\big)=\iota_{\{0\}}(u-Ax)=\iota_{\operatorname{gra}A}(x,u)\\&=\iota_{\operatorname{gra}A}(x,u)+\langle x,Ax\rangle=F_A^*(u,x),\end{aligned} \tag{17}$$

which completes the proof. ∎

We now present a new characterization of skew operators using the Fitzpatrick family.

THEOREM 2.9. *Let $A\colon X\to X$ be continuous, linear, and monotone. Then A is skew $\Leftrightarrow$ $\mathcal{F}_A$ is a singleton. In this case, $\mathcal{F}_A=\{\iota_{\operatorname{gra}A}\}$.*

Proof. Fix $(x,u)\in X\times X$.

"$\Leftarrow$": If $u-Ax\notin\operatorname{ran}A_+$, then $u-Ax\neq 0$. Now suppose that $u-Ax\neq 0$. Then $(x,u)\notin\operatorname{gra}A$ and hence $F_A^*(u,x)=+\infty$ by Theorem 2.3(iii). Fact 1.6 implies that $F_A(x,u)=+\infty$, i.e., $(x,u)\notin\operatorname{dom}F_A$. If $u+A^*x$ belonged to $\operatorname{ran}A_+$, then $q_{A_+}^*(u+A^*x)<+\infty$ (by Fact 2.2(iii)) and hence $(x,u)\in\operatorname{dom}F_A$ (by Theorem 2.3(i)), which is absurd. Thus $u+A^*x\notin\operatorname{ran}A_+$. Now $u+A^*x=u-Ax+2A_+x$, which implies $u-Ax\notin\operatorname{ran}A_+$. Altogether, we have verified the equivalence

$$\big(\forall(x,u)\in X\times X\big)\quad u-Ax\neq 0 \;\Leftrightarrow\; u-Ax\notin\operatorname{ran}A_+. \tag{18}$$

Since $(\forall u\in X\smallsetminus\{0\})\ u-A0=u\neq 0$, (18) yields $u=u-A0\notin\operatorname{ran}A_+$. Hence $\operatorname{ran}A_+=\{0\}$; equivalently, $A_+=0$ and therefore $A=A_\circ$.

"$\Rightarrow$": This follows from Example 2.8 and Fact 1.6. ∎

Remark 2.10. Loosely speaking, Theorem 2.9 states that a Fitzpatrick family with only one element corresponds to a "bad" (here, skew) monotone operator. The situation is similar for subdifferential operators: $\mathcal{F}_{\partial f}$ reduces to the singleton $\{f\oplus f^*\}$ when f is sublinear or an indicator function (see [12, 13] and also [2]).

3. Range. In this section, we compare the range of a continuous linear monotone operator to the range of its adjoint.

PROPOSITION 3.1. *Let $A: X \to X$ be continuous, linear, and monotone. Then* $\ker A = \ker A^*$ *and* $\overline{\operatorname{ran}}\, A = \overline{\operatorname{ran}}\, A^*$.

Proof. Take $x \in \ker A$ and $v \in \operatorname{ran} A$, say $v = Ay$. Then $(\forall \alpha \in \mathbb{R})\ 0 \leq \langle \alpha x + y, A(\alpha x + y)\rangle = \alpha\langle x, v\rangle + \langle y, Ay\rangle$. Hence $\langle x, v\rangle = 0$ and thus $\ker A \subset (\operatorname{ran} A)^\perp = \ker A^*$. Since A^* is also continuous, linear, and monotone, we obtain $\ker A^* \subset \ker A^{**} = \ker A$. Altogether, $\ker A = \ker A^*$ and therefore $\overline{\operatorname{ran}}\, A = \overline{\operatorname{ran}}\, A^*$. ∎

Remark 3.2.

(i) A referee pointed out that Proposition 3.1 also follows from [30, Corollary 3.5], which is more general.

(ii) Example 3.3 below illustrates that the closures in Proposition 3.1 are critical.

(iii) An operator $A: X \to X$ such that $\operatorname{ran} A = \operatorname{ran} A^*$ is called *range-symmetric* or *EP*; see [26, p. 408]. Proposition 3.1 implies that every continuous, linear, and monotone operator with closed range is range-symmetric. See [16, Theorem 2.3] for equivalent properties in the matrix case.

(iv) Every *normal matrix* A (i.e., $AA^* = A^*A$) is range-symmetric: indeed, we then have $\operatorname{ran} A = \operatorname{ran} AA^* = \operatorname{ran} A^*A = \operatorname{ran} A^*$ (the first and the last equalities follow, e.g., from [26, p. 212]).

(v) However, a range-symmetric monotone matrix need not be normal:

$$A = \begin{pmatrix} 2 & 1 \\ -1 & 1 \end{pmatrix} \tag{19}$$

is monotone, but $AA^* \neq A^*A$.

Example 3.3 (Volterra operator). Set $X = L_2[0,1]$. The *Volterra integration operator* [21, Problem 148] is defined by

$$V: X \to X: x \mapsto Vx, \quad \text{where} \quad Vx: [0,1] \to \mathbb{R}: t \mapsto \int_0^t x. \tag{20}$$

Fix $x \in X$. Then

$$(V^*x)(t) = \int_t^1 x \tag{21}$$

and $\ker V = \ker V^* = \{0\}$, so V and V^* have dense range. Set $e \equiv 1 \in X$. Now (20) and (21) imply $(V + V^*)x = \langle x, e\rangle e$ and thus $\langle x, (V + V^*)x\rangle = \langle x, e\rangle^2 \geq 0$. Hence

$$V \text{ is monotone and } V_+x = \tfrac{1}{2}\langle x, e\rangle e. \tag{22}$$

Moreover, $q_{V_+}(x) = \frac{1}{2}\langle x, V_+x\rangle = \frac{1}{4}\langle x, e\rangle^2$ and $\operatorname{ran} V_+ = \mathbb{R}e$ is closed. Now Fact 2.2(ii) and Theorem 2.3(i), (iii) result in

$$F_V: X \times X \to\ ]-\infty, +\infty] \tag{23}$$
$$(z, w) \mapsto \begin{cases} \frac{1}{2}\langle w + V^*z, e\rangle^2 & \text{if } w + V^*z = \langle w + V^*z, e\rangle e; \\ +\infty & \text{otherwise} \end{cases}$$

and

$$F_V^*: X \times X \to\ ]-\infty, +\infty] \tag{24}$$
$$(w, z) \mapsto \begin{cases} \frac{1}{2}\langle z, e\rangle^2 & \text{if } w = Vz; \\ +\infty & \text{otherwise.} \end{cases}$$

Next, assume that $Vx = V^*y$, i.e., $(\forall t \in [0,1])\ \int_0^t x = \int_t^1 y$. Evaluating this at $t = 0$ and $t = 1$, we learn that $\langle y, e\rangle = \langle x, e\rangle = 0$. We thus have verified the implication

$$\left.\begin{array}{c} x \in X \\ y \in X \\ Vx = V^*y \end{array}\right\} \quad \Rightarrow \quad \langle x, e\rangle = \langle y, e\rangle = 0 \tag{25}$$

and the inclusion

$$\operatorname{ran} V \cap \operatorname{ran} V^* \subseteq \big\{Vx \mid x \in \{e\}^\perp\big\}. \tag{26}$$

Conversely, if $x \in \{e\}^\perp$, then

$$(\forall t \in [0,1]) \quad (Vx)(t) = \langle x, e\rangle - \int_t^1 x = \big(V^*(-x)\big)(t) \tag{27}$$

and hence $Vx \in \operatorname{ran} V \cap \operatorname{ran} V^*$. Altogether,

$$\operatorname{ran} V \cap \operatorname{ran} V^* = \{Vx \colon x \in \{e\}^\perp\}. \tag{28}$$

Since $\langle e, e\rangle = 1 \neq 0$, the implication (25) shows that $Ve \notin \operatorname{ran} V^*$ and that $V^*e \notin \operatorname{ran} V$. Therefore,

$$\operatorname{ran} V \nsubseteq \operatorname{ran} V^* \ \text{ and } \ \operatorname{ran} V^* \nsubseteq \operatorname{ran} V. \tag{29}$$

4. Rectangular monotone operators. We now turn to a property related to the domain of the Fitzpatrick function.

DEFINITION 4.1 (rectangular). *Let $A \colon X \to X$ be continuous, linear, and monotone. Then A is* rectangular *if $X \times \operatorname{ran} A \subseteq \operatorname{dom} F_A$.*

Remark 4.2.

(i) The property referred to in Definition 4.1 was first introduced by Brézis and Haraux [11]. In the literature it is also known as *property (*)* and as *3*-monotone*. However, we follow here Simons' [39] more descriptive naming convention, which is based on his observation that—since $\operatorname{dom} F_A \subseteq \overline{\operatorname{dom}}\, A \times \overline{\operatorname{ran}}\, A = X \times \overline{\operatorname{ran}}\, A$ is always true—the operator A is rectangular if and only if $\overline{\operatorname{dom}}\, F_A$ is the "rectangle" $X \times \overline{\operatorname{ran}}\, A$.

(ii) In the context of general monotone operators, the subdifferential operator is known to be rectangular [11].

(iii) As a consequence of (ii), we note that every continuous, linear, monotone, and symmetric operator is rectangular (Fact 2.2(i)). This will be reproved in Corollary 4.9 below.

The importance of rectangularity stems from a powerful result due to Brézis and Haraux [11], which we state next in the present context of linear operators.

FACT 4.3 (Brézis–Haraux). *Let A and B be continuous, linear, and monotone operators from X to X, and suppose that A or B is rectangular. Then $\overline{\operatorname{ran}}\,(A+B) = \overline{\operatorname{ran} A + \operatorname{ran} B}$ and $\operatorname{int} \operatorname{ran}(A+B) = \operatorname{int}(\operatorname{ran} A + \operatorname{ran} B)$.*

Proof. See [11], and also [36, 39] for different proofs. ∎

It is worthwhile to list some of the most important consequences of Fact 4.3.

COROLLARY 4.4. *Let A and B be continuous, linear, and monotone operators from X to X. Suppose that A or B is rectangular, and that A or B is surjective. Then $A + B$ is surjective.*

Proof. Fact 4.3 yields $X = \operatorname{int} X = \operatorname{int}(\operatorname{ran} A + \operatorname{ran} B) = \operatorname{int} \operatorname{ran}(A+B)$. Therefore, $X = \operatorname{ran}(A+B)$ and $A+B$ is surjective. ∎

COROLLARY 4.5. *Let A and B be continuous, linear, and monotone operators from X to X such that A or B is rectangular. Then* $\ker(A+B) = \ker A \cap \ker B$.

Proof. Using Proposition 3.1 and Fact 4.3, we obtain

$$\begin{aligned}(\ker A \cap \ker B)^\perp &= \overline{(\ker A)^\perp + (\ker B)^\perp} = \overline{\operatorname{ran} A^* + \operatorname{ran} B^*} \\ &= \overline{\operatorname{ran} A + \operatorname{ran} B} = \overline{\operatorname{ran}}\,(A+B) \\ &= \big(\ker(A+B)\big)^\perp. \end{aligned} \tag{30}$$

The result follows by taking orthogonal complements. ∎

COROLLARY 4.6. *Let A and B be continuous, linear, and monotone operators from X to X. Suppose that A or B is rectangular, and that A or B is injective. Then $A+B$ is injective.*

COROLLARY 4.7. *Let A and B be continuous, linear, and monotone operators from X to X. Suppose that A or B is rectangular, and that A or B is bijective. Then $A+B$ is bijective.*

PROPOSITION 4.8. *Let $A\colon X \to X$ be continuous, linear, and monotone. Then the following are equivalent:*

(i) *A is rectangular.*
(ii) $\operatorname{ran} A + \operatorname{ran} A^* \subseteq \operatorname{dom} q^*_{A_+}$.
(iii) $\operatorname{ran} A_\circ \subseteq \operatorname{dom} q^*_{A_+}$.

Proof. (i)⇔(ii): This is a direct consequence of Theorem 2.3(i). (ii)⇒(iii): $\operatorname{ran} A_\circ = \operatorname{ran}(A - A^*) \subseteq \operatorname{ran} A - \operatorname{ran} A^* = \operatorname{ran} A + \operatorname{ran} A^* \subseteq \operatorname{dom} q^*_{A_+}$. (ii)⇐(iii): This follows from Fact 2.2(iii), (iv) and the fact that $A^* = A_+ - A_\circ$ yield $\operatorname{ran} A + \operatorname{ran} A^* = \operatorname{ran}(A_+ + A_\circ) + \operatorname{ran}(A_+ - A_\circ) \subseteq \operatorname{ran} A_+ + \operatorname{ran} A_\circ \subseteq \operatorname{dom} q^*_{A_+} + \operatorname{dom} q^*_{A_+} = \operatorname{dom} q^*_{A_+}$. ∎

COROLLARY 4.9. *Let $A\colon X \to X$ be continuous, linear, monotone, and symmetric. Then A is rectangular.*

Proof. Utilizing Fact 2.2(iii), we see that $\operatorname{ran} A + \operatorname{ran} A^* = \operatorname{ran} A_+ \subseteq \operatorname{dom} q^*_{A_+}$. The result follows from Proposition 4.8. ∎

COROLLARY 4.10. *Let $A\colon X \to X$ be continuous, linear, and monotone, and suppose that* $\operatorname{ran} A_+$ *is closed. Then A is rectangular if and only if* $\operatorname{ran} A_\circ \subseteq \operatorname{ran} A_+$.

Proof. Fact 2.2(iii) shows that $\operatorname{dom} q^*_{A_+} = \operatorname{ran} A_+$. Applying Proposition 4.8, we obtain the proof. ∎

Remark 4.11 (paramonotone operators). Let $X = \mathbb{R}^n$ and let $A \in \mathbb{R}^{n\times n}$ be monotone. By [22, Proposition 3.2.(ii)], A is *paramonotone* ⇔ $\ker A_+ \subseteq \ker A$. On the other hand, using Corollary 4.5 (applied to A_+ and $A_\circ$) and Corollary 4.10, we have the equivalences $\ker A_+ \subseteq \ker A \Leftrightarrow \ker A_+ \subseteq \ker A_+ \cap \ker A_\circ \Leftrightarrow \ker A_+ \subseteq \ker A_\circ$ ⇔ $\operatorname{ran} A_\circ \subseteq \operatorname{ran} A_+ \Leftrightarrow A$ is rectangular. Altogether,

A is paramonotone if and only if A is rectangular. (31)

See [22] for further information on paramonotone operators.

The next result can be deduced from [11, Proposition 2]. The proof provided here is somewhat simpler and based on the Fitzpatrick function, and the result is stated in a more applicable form.

THEOREM 4.12. *Let $A\colon X \to X$ be continuous, linear, and monotone. Then the following are equivalent:*

(i) *A is rectangular.*
(ii) *For some $\gamma > 0$, $\|\gamma A - \mathrm{Id}\| \leq 1$.*
(iii) *A^* is rectangular.*

Proof. The conditions all hold when $A = 0$, so assume that $A \neq 0$.

(i)⇒(ii): Consider the function

$$f\colon X \to \left]-\infty,+\infty\right]\colon x \mapsto F_A(x,0). \tag{32}$$

Then f is convex, lower semicontinuous, and proper by Fact 1.4(i), (ii). Since A is rectangular, $X \times \{0\} \subseteq X \times \operatorname{ran} A \subseteq \operatorname{dom} F_A$. Hence $\operatorname{dom} f = X$. It follows, e.g., from [43, Theorem 2.2.20] that there exist $\delta > 0$ and $\beta > 0$ such that $(\forall x \in B(0;\delta))$ $f(x) = F_A(x,0) = \sup_{y\in X}\langle x, Ay\rangle - \langle y, Ay\rangle \leq \beta$. Fix $x \in B(0;\delta)$ and $y \in X$. Then

$$(\forall \rho \in \mathbb{R}) \quad 0 \leq \beta + \langle \rho y, A(\rho y)\rangle - \langle x, A(\rho y)\rangle = \beta + \rho^2\langle y, Ay\rangle - \rho\langle x, Ay\rangle. \tag{33}$$

We claim that

$$(\forall x \in B(0;\delta))(\forall y \in X) \quad \langle x, Ay\rangle^2 \leq 4\beta\langle y, Ay\rangle. \tag{34}$$

If $\langle y, Ay\rangle = 0$, then (33) shows that $\langle x, Ay\rangle = 0$, and hence (34) holds. Now assume that $\langle y, Ay\rangle \neq 0$. In terms of ρ, the right side of (33) is a nonnegative quadratic function. Substituting the minimizer $\langle x, Ay\rangle/(2\langle y, Ay\rangle)$ of this quadratic function into (33) yields an inequality that is equivalent to (34). In turn, (34) leads to

$$(\forall y \in X) \quad \delta^2\|Ay\|^2 \leq 4\beta\langle y, Ay\rangle. \tag{35}$$

Set $\alpha = \delta^2/(4\beta)$. We deduce that $(\forall y \in X)$ $\langle y, \alpha Ay\rangle \geq \|\alpha Ay\|^2$, i.e., αA is firmly nonexpansive. This (see [19]) is equivalent to the nonexpansivity of $2\alpha A - \mathrm{Id}$, i.e., to $\|2\alpha A - \mathrm{Id}\| \leq 1$.

(ii)⇒(i): Set $\alpha = \gamma/2$. Fix x and y in X and take $z \in X$. Utilizing the equivalences αA is firmly nonexpansive $\Leftrightarrow \|2\alpha A - \mathrm{Id}\| \leq 1 \Leftrightarrow \|2\alpha A^* - \mathrm{Id}\| \leq 1 \Leftrightarrow \alpha A^*$ is firmly nonexpansive, we estimate

$$\begin{aligned}\langle x, Az\rangle + \langle z, Ay\rangle - \langle z, Az\rangle &= \big(\langle x, Az\rangle - \tfrac{1}{2}\langle z, Az\rangle\big) + \big(\langle A^*z, y\rangle - \tfrac{1}{2}\langle z, A^*z\rangle\big)\\ &\leq \big(\|x\|\,\|Az\| - \tfrac{1}{2}\alpha\|Az\|^2\big) + \big(\|A^*z\|\,\|y\| - \tfrac{1}{2}\alpha\|A^*z\|^2\big)\\ &\leq \tfrac{1}{2\alpha}\big(\|x\|^2 + \|y\|^2\big),\end{aligned} \tag{36}$$

where the last inequality was obtained by computing the maxima of the quadratic functions $\rho \mapsto \|x\|\rho - \frac{1}{2}\alpha\rho^2$ and $\rho \mapsto \|y\|\rho - \frac{1}{2}\alpha\rho^2$, respectively. It follows from (36) that

$$(\forall x \in X)(\forall y \in X) \quad F_A(x, Ay) \leq \tfrac{1}{\gamma}\big(\|x\|^2 + \|y\|^2\big), \tag{37}$$

hence $X \times \operatorname{ran} A \subset \operatorname{dom} F_A$.

(ii)⇔(iii): Apply the equivalence (i)⇔(ii) to A^*. □

COROLLARY 4.13. *The continuous, linear, monotone, and rectangular operators form a convex cone.*

Proof. It is clear that they form a cone. Suppose A and B are continuous, linear, monotone, and rectangular. Then there exist $\gamma_A > 0$ and $\gamma_B > 0$ such that $\|\gamma_A A - \mathrm{Id}\| \leq 1$ and $\|\gamma_B B - \mathrm{Id}\| \leq 1$. Set $\gamma = \frac{1}{2}\min\{\gamma_A, \gamma_B\}$ and estimate

$\|\gamma(A+B)-\mathrm{Id}\| \leq \frac{1}{2}\|2\gamma A-\mathrm{Id}\|+\frac{1}{2}\|2\gamma B-\mathrm{Id}\| \leq 1$. Hence $A+B$ is rectangular and the proof is complete. ∎

The next example was established by direct computation in [4]; however, Theorem 4.12 yields a very transparent and simple proof.

Example 4.14. Let $R\colon X^n \to X^n\colon (x_1,x_2,\ldots,x_n) \mapsto (x_n,x_1,\ldots,x_{n-1})$ be the *right-shift operator* on X^n. Then $\mathrm{Id}-R$ is rectangular.

Proof. Since $\|1\cdot(\mathrm{Id}-R)-\mathrm{Id}\| = \|-R\| = 1$, the result is a consequence of Theorem 4.12. ∎

We conclude this section by providing a novel nonsmooth proof of a result on the domain of the Fitzpatrick function of the subdifferential operator (see also [6, Theorem 2.6]).

THEOREM 4.15. *Let $f\colon X \to]-\infty,+\infty]$ be convex, lower semicontinuous, and proper. Then*

$$\operatorname{dom} f \times \operatorname{dom} f^* \subseteq \operatorname{dom} F_{\partial f} \subseteq \overline{\operatorname{dom}}\, f \times \overline{\operatorname{dom}}\, f^*. \tag{38}$$

Proof. The first inclusion is elementary (see also [6, Proposition 2.1]). Now take $(x,u) \in \operatorname{dom} F_{\partial f}$ and set $C = \overline{\operatorname{dom}}\, f$. Assume to the contrary that $x \notin C$; hence $f(x) = +\infty$ and $d_C(x) = \inf\|x-C\| > 0$. Fix $x_0 \in \operatorname{dom} f$ and define the family of nonconvex but lower semicontinuous functions

$$(\forall \rho > 0) \quad f_\rho\colon X \to]-\infty,+\infty]\colon y \mapsto \begin{cases} f(y) & \text{if } y \neq x;\\ f(x_0)+\rho & \text{if } y = x. \end{cases} \tag{39}$$

The *approximate mean value theorem* of Mordukhovich and Shao (see [27, Theorem 3.49] or [28, Theorem 8.2]), applied to f_ρ and the points x_0 and x, shows that for every $\rho > 0$, there exist $y_\rho \in [x_0,x[$ and a sequence $(y_{\rho,n},v_{\rho,n})_{n\in\mathbb{N}}$ in $\operatorname{gra}\partial f$ such that $y_{\rho,n} \to y_\rho$ and

$$\varliminf_{n\in\mathbb{N}} \Big\langle \frac{x-y_{\rho,n}}{\|x-y_{\rho,n}\|}, v_{\rho,n}\Big\rangle \geq \frac{f_\rho(x)-f_\rho(x_0)}{\|x-x_0\|} = \frac{\rho}{\|x-x_0\|}. \tag{40}$$

Therefore, there exists a sequence $((z_n,w_n))_{n\in\mathbb{N}}$ in $\operatorname{gra}\partial f$ such that

$$\Big\langle \frac{x-z_n}{\|x-z_n\|}, w_n\Big\rangle \to +\infty. \tag{41}$$

By definition of $F_{\partial f}$, the Cauchy–Schwarz inequality, and (41), we obtain

$$\begin{aligned} F_{\partial f}(x,u) &= \sup_{(y,v)\in\operatorname{gra}\partial f} \big(\langle x,v\rangle + \langle y,u\rangle - \langle y,v\rangle\big)\\ &= \sup_{(y,v)\in\operatorname{gra}\partial f} \big(\langle x-y,v\rangle + \langle y-x,u\rangle + \langle x,u\rangle\big)\\ &\geq \sup_{(y,v)\in\operatorname{gra}\partial f} \left(\|x-y\|\Big(\Big\langle \frac{x-y}{\|x-y\|}, v\Big\rangle - \|u\|\Big) + \langle x,u\rangle\right)\\ &\geq \varlimsup_{n\in\mathbb{N}} \left(\|x-z_n\|\Big(\Big\langle \frac{x-z_n}{\|x-z_n\|}, w_n\Big\rangle - \|u\|\Big) + \langle x,u\rangle\right)\\ &\geq \varlimsup_{n\in\mathbb{N}} \left(d_C(x)\Big(\Big\langle \frac{x-z_n}{\|x-z_n\|}, w_n\Big\rangle - \|u\|\Big) + \langle x,u\rangle\right)\\ &= +\infty. \end{aligned} \tag{42}$$

This contradicts the assumption that $F_{\partial f}(x,u) < +\infty$. Therefore, $x \in \overline{\operatorname{dom}}\, f$. An analogous argument (applied to f^*) implies that $u \in \overline{\operatorname{dom}}\, f^*$. ∎

5. The Fitzpatrick function of the sum. One of Fitzpatrick's open problems [17, Problem 5.4] is to find the Fitzpatrick function of the sum of two operators. This has proven to be a difficult problem. However, an upper bound is always readily available.

DEFINITION 5.1. *Let $A\colon X \to X$ and $B\colon X \to X$ be continuous, linear, and monotone operators, and set*

$$(\forall (x,u) \in X \times X) \quad \Phi_{\{A,B\}}(x,u) = \big(F_A(x,\cdot) \,\square\, F_B(x,\cdot)\big)(u) = \inf_{v+w=u} F_A(x,v) + F_B(x,w). \tag{43}$$

PROPOSITION 5.2 (upper bound). *Let $A\colon X \to X$ and $B\colon X \to X$ be continuous, linear, and monotone operators. Then $F_{A+B} \leq \Phi_{\{A,B\}}$.*

Proof. See [6, Proposition 4.2]. □

In [6, section 4] it is shown that in the context of subdifferential operators, this upper bound is sometimes—but not always—tight. In the remainder of this section we investigate the upper bound in the present context of continuous, linear, and monotone operators.

THEOREM 5.3. *Let $A\colon X \to X$ and $B\colon X \to X$ be continuous, linear, and monotone operators. Suppose that one of the following conditions is satisfied:*

(i) *A is skew, and B is skew.*

(ii) *A is symmetric, and B is skew.*

Then $F_{A+B} = \Phi_{\{A,B\}}$.

Proof. Fix $(x,u) \in X \times X$. (i) Repeated application of Example 2.8 yields

$$\begin{aligned} F_{A+B}(x,u) &= \iota_{\operatorname{gra}(A+B)}(x,u) \\ &= \inf_{v+w=u} \iota_{\{Ax\}}(v) + \iota_{\{Bx\}}(w) \\ &= \inf_{v+w=u} \iota_{\operatorname{gra} A}(x,v) + \iota_{\operatorname{gra} B}(x,w) \\ &= \inf_{v+w=u} F_A(x,v) + F_B(x,w) \\ &= \Phi_{\{A,B\}}(x,u). \end{aligned} \tag{44}$$

(ii) Theorem 2.3(i) and Example 2.8 result in

$$\begin{aligned} F_{A+B}(x,u) &= F_A(x, u - Bx) \\ &= \inf_{v \in Bx} F_A(x, u - v) \\ &= \inf_{v \in X} F_A(x, u - v) + \iota_{\operatorname{gra} B}(x,v) \\ &= \inf_{v \in X} F_A(x, u - v) + F_B(x,v) \\ &= \Phi_{\{A,B\}}(x,u). \end{aligned} \tag{45}$$

The proof is complete. □

The "purely symmetric" counterpart to Theorem 5.3 seems to require a closedness assumption. We are grateful to a referee for providing us with a simpler and more powerful proof.

THEOREM 5.4. *Let $A\colon X \to X$ and $B\colon X \to X$ be continuous, linear, monotone, and symmetric. Then $F_{A+B} = \Phi_{\{A,B\}}$ on $\operatorname{ran}(A+B)$. Consequently, if $\operatorname{ran}(A+B)$ is closed, then $F_{A+B} = \Phi_{\{A,B\}}$.*

Proof. Fix $x \in X$ and $y \in X$. Utilizing Example 2.4 thrice, we obtain

$$\begin{aligned}\Phi_{\{A,B\}}\big(x,(A+B)y\big) &\leq F_A(x,Ay)+F_B(x,By)\\ &= \tfrac{1}{4}\langle x+y, A(x+y)\rangle + \tfrac{1}{4}\langle x+y, B(x+y)\rangle\\ &= \tfrac{1}{4}\langle x+y,(A+B)(x+y)\rangle\\ &= F_{A+B}\big(x,(A+B)y\big).\end{aligned} \tag{46}$$

Thus

$$\Phi_{\{A,B\}} \leq F_{A+B} \text{ on } X\times \operatorname{ran}(A+B). \tag{47}$$

On the other hand, by Proposition 5.2, $F_{A+B} \leq \Phi_{\{A,B\}}$. Altogether, $F_{A+B} = \Phi_{\{A,B\}}$ on $X\times\operatorname{ran}(A+B)$. If $\operatorname{ran}(A+B)$ is closed, we deduce from (12) that $F_{A+B} = \Phi_{\{A,B\}}$ everywhere. ∎

Remark 5.5. We do not know whether or not the conclusion of Theorem 5.4 remains true when the assumption on the closedness of the range of the sum of the operators is omitted. Indeed, we do not know whether or not two continuous, linear, and monotone operators $A\colon X\to X$ and $B\colon X\to X$ exist for which $F_{A+B}\neq\Phi_{\{A,B\}}$.

COROLLARY 5.6. *Let $A\colon X\to X$ and $B\colon X\to X$ be continuous, linear, and monotone operators such that* $\operatorname{ran}(A_+ + B_+)$ *is closed. Then $F_{A+B} = \Phi_{\{A,B\}}$.*

Proof. Fix $(x,u)\in X\times X$. Using Theorems 5.3(ii), 5.4, 5.3(i), and 5.3(ii) again, we obtain

$$\begin{aligned}&F_{A+B}(x,u)\\ &= F_{A_+ + A_\circ + B_+ + B_\circ}(x,u) = F_{(A_+ + B_+)+(A_\circ+B_\circ)}(x,u)\\ &= \inf_{v+w=u} F_{A_+ + B_+}(x,v) + F_{A_\circ+B_\circ}(x,w)\\ &= \inf_{v+w=u}\Big(\inf_{v_1+v_2=v} F_{A_+}(x,v_1)+F_{B_+}(x,v_2) + \inf_{w_1+w_2=w} F_{A_\circ}(x,w_1)+F_{B_\circ}(x,w_2)\Big)\\ &= \inf_{v_1+v_2+w_1+w_2=u} F_{A_+}(x,v_1)+F_{A_\circ}(x,w_1)+F_{B_+}(x,v_2)+F_{B_\circ}(x,w_2)\\ &= \inf_{u_1+u_2=u}\Big(\inf_{v_1+w_1=u_1} F_{A_+}(x,v_1)+F_{A_\circ}(x,w_1) + \inf_{v_2+w=u_2} F_{B_+}(x,v_2)+F_{B_\circ}(x,w_2)\Big)\\ &= \inf_{u_1+u_2=u} F_A(x,u_1)+F_B(x,u_2)\\ &= \Phi_{\{A,B\}}(x,u),\end{aligned} \tag{48}$$

as required. ∎

COROLLARY 5.7. *Suppose that X is finite-dimensional, and let $A\colon X\to X$ and $B\colon X\to X$ be continuous, linear, and monotone operators. Then $F_{A+B} = \Phi_{\{A,B\}}$.*

6. Cyclic monotonicity. An interesting quantitative grading of monotonicity is the notion of cyclic monotonicity of order n. As demonstrated in [2], this property is captured with a Fitzpatrick function of the corresponding order. In this section, we study these notions for continuous linear operators. Let us start with the relevant definitions.

DEFINITION 6.1 (n-cyclic monotonicity). *Let $A\colon X\to X$ be continuous and linear. Then A is* n-cyclically monotone *if $n\in\{2,3,\dots\}$ and*

$$(\forall (x_1,\dots,x_n)\in X^n)\quad \left(\sum_{i=1}^{n-1}\langle x_{i+1}-x_i, Ax_i\rangle\right) + \langle x_1-x_n, Ax_n\rangle \leq 0. \tag{49}$$

The operator A is cyclically monotone *if A is n-cyclically monotone for every $n \in \{2,3,\dots\}$.*

Note that an operator is monotone if and only if it is 2-cyclically monotone.

DEFINITION 6.2 (Fitzpatrick function of order n). *Let $A\colon X \to X$. For every $n \in \{2,3,\dots\}$, the* Fitzpatrick function of A of order n *is*

(50)
$$F_{A,n}(x,u) = \sup_{(x_1,\dots,x_{n-1})\in X^{n-1}} \Big(\sum_{i=1}^{n-2}\langle x_{i+1}-x_i, Ax_i\rangle\Big) + \langle x - x_{n-1}, Ax_{n-1}\rangle + \langle x_1, u\rangle.$$

We set $F_{A,\infty} = \sup_{n\in\{2,3,\dots\}} F_{A,n}$.

Note that $F_{A,2} = F_A$. We refer the reader to [2], where it is shown that $F_{A,n}$ is well suited to study n-cyclic monotonicity of A. Most relevant for our current setting is the following result.

FACT 6.3 (see [2, Theorem 2.9]). *Let $A\colon X \to X$ be maximal monotone, and let $n \in \{2,3,\dots\}$. Then A is n-cyclically monotone* $\Leftrightarrow \operatorname{gra} A = \big\{(x,u) \in X\times X \mid F_{A,n}(x,u) = \langle x,u\rangle\big\}$.

Let us compute the Fitzpatrick functions of an arbitrary continuous, linear, symmetric, and positive definite operator. This result generalizes [2, Example 4.4].

Example 6.4. Let $A\colon X \to X$ be continuous, linear, symmetric, and positive definite, and let $n \in \{2,3,\dots\}$. Then

$$F_{A,n}\colon X\times X \to \mathbb{R}\colon (x,u) \mapsto \tfrac{n-1}{2n}\big(\|x\|_A^2 + \|u\|_{A^{-1}}^2\big) + \tfrac{1}{n}\langle x,u\rangle \tag{51}$$

and

$$F_{A,\infty} = \tfrac{1}{2}\|\cdot\|_A^2 \oplus \tfrac{1}{2}\|\cdot\|_{A^{-1}}^2. \tag{52}$$

Proof. By [2, Example 4.4], we have

$$F_{\mathrm{Id},n}\colon X\times X \to \mathbb{R}\colon (x,u) \mapsto \tfrac{n-1}{2n}\big(\|x\|^2 + \|u\|^2\big) + \tfrac{1}{n}\langle x,u\rangle \tag{53}$$

and

$$F_{\mathrm{Id},\infty} = \tfrac{1}{2}\|\cdot\|^2 \oplus \tfrac{1}{2}\|\cdot\|^2. \tag{54}$$

Fix $(x,u) \in X\times X$. By definition, $F_{A,n}(x,u)$ is equal to

$$\begin{aligned}&\sup_{(x_1,\dots,x_{n-1})\in X^{n-1}} \Big(\sum_{i=1}^{n-2}\langle x_{i+1}-x_i, Ax_i\rangle\Big) + \langle x - x_{n-1}, Ax_{n-1}\rangle + \langle x_1,u\rangle \\ &= \sup_{(x_1,\dots,x_{n-1})\in X^{n-1}} \Big(\sum_{i=1}^{n-2}\langle x_{i+1}-x_i, x_i\rangle_A\Big) + \langle x - x_{n-1}, x_{n-1}\rangle_A + \langle x_1, A^{-1}u\rangle_A.\end{aligned} \tag{55}$$

The result now follows by applying (53) and (54) to Id, viewed as an operator on $(X, \langle\cdot,\cdot\rangle_A)$. ∎

We now provide a simple, yet powerful, recursion formula.

THEOREM 6.5 (recursion). *Let $A\colon X \to X$ be monotone, and let $n \in \{2,3,\dots\}$. Then*

$$\big(\forall (x,u)\in X\times X\big)\quad F_{A,n+1}(x,u) = \sup_{y\in X}\big(F_{A,n}(y,u) + \langle x-y, Ay\rangle\big). \tag{56}$$

Proof. Fix $(x,u) \in X \times X$. Using the definition, we see that $F_{A,n+1}(x,u)$ is equal to

$$\begin{aligned}
&\sup_{(x_1,\ldots,x_n)\in X^n} \Big(\sum_{i=1}^{n-1} \langle x_{i+1}-x_i, Ax_i\rangle \Big) + \langle x-x_n, Ax_n\rangle + \langle x_1, u\rangle \\
&= \sup_{x_n\in X} \Bigg(\sup_{(x_1,\ldots,x_{n-1})\in X^{n-1}} \Big(\sum_{i=1}^{n-2} \langle x_{i+1}-x_i, Ax_i\rangle \Big) + \langle x_n - x_{n-1}, Ax_{n-1}\rangle + \langle x_1, u\rangle \Bigg) \\
&\qquad\qquad + \langle x-x_n, Ax_n\rangle \\
&= \sup_{x_n\in X} \Big(F_{A,n}(x_n,u) + \langle x-x_n, Ax_n\rangle \Big).
\end{aligned} \tag{57}$$

The proof is complete. ∎

This section is concluded with two results on the domain of the Fitzpatrick function of order n.

THEOREM 6.6. *Let $f\colon X \to \left]-\infty,+\infty\right]$ be convex, lower semicontinuous, and proper, and let $n \in \{2,3,\ldots\}$. Then*

$$\operatorname{dom} f \times \operatorname{dom} f^* \subseteq \operatorname{dom} F_{\partial f,n} \subseteq \operatorname{dom} F_{\partial f} \subseteq \overline{\operatorname{dom}}\, f \times \overline{\operatorname{dom}}\, f^*. \tag{58}$$

Proof. By [2, Theorem 3.5], we know that $F_{\partial f,n} \le f \oplus f^*$, which implies the first inequality of (58). The second inequality is clear since $(F_{\partial f,n})_{n\in\{2,3,\ldots\}}$ is an increasing sequence. The third inequality follows from Theorem 4.15. ∎

COROLLARY 6.7. *Let $A\colon X \to X$ be continuous, linear, monotone, and symmetric, and let $n \in \{2,3,\ldots\}$. Then*

$$X \times \operatorname{ran} A \subseteq \operatorname{dom} F_{A,n} \subseteq X \times \overline{\operatorname{ran}}\, A. \tag{59}$$

7. Rotators in the Euclidean plane. This section covers rotators in the Euclidean plane. We characterize their cyclic monotonicity properties, and we provide formulas for the Fitzpatrick function of any order.

From now on, $X = \mathbb{R}^2$ and

$$A_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \quad \text{where } \theta \in [0,\pi/2]. \tag{60}$$

The main result of this section will be stated at the end. For clarity of presentation, we break up the proof into several propositions. The first proposition characterizes n-cyclic monotonicity. See also Asplund's paper [1] for characterizations for general matrices.

PROPOSITION 7.1. *Let $n \in \{2,3,\ldots\}$. Then A_θ is n-cyclically monotone $\Leftrightarrow$ $\theta \in [0,\pi/n]$.*

Proof. If $n = 2$, then the symmetric part of A_θ is $\cos\theta\,\mathrm{Id}$ and the equivalence is clear. Thus, we assume that $n \in \{3,4,\ldots\}$. We shall characterize the n-cyclic monotonicity of A_θ in terms of the positive semidefiniteness of an associated Hermitian matrix. Take n points $x_1 = (\xi_1,\eta_1),\ldots,x_n = (\xi_n,\eta_n)$ in X, and set $x_{n+1} = x_1$. We must show that

$$0 \ge \sum_{i=1}^{n} \langle x_{i+1}-x_i, A_\theta x_i\rangle. \tag{61}$$

Let us identify $\mathbb{R}^2$ with $\mathbb{C}$ in the standard way: $x = (\xi, \eta)$ in $\mathbb{R}^2$ corresponds to $\xi + \mathrm{i}\eta$ in $\mathbb{C}$, where $\mathrm{i} = \sqrt{-1}$, and $\langle x, y\rangle = \mathrm{Re}(\overline{x}y)$ for x and y in $\mathbb{C}$. The operator A_θ corresponds to complex multiplication by

$$\omega = \exp(\mathrm{i}\theta). \tag{62}$$

Thus we aim to show that $0 \geq \mathrm{Re}\left(\sum_{i=1}^n (\overline{x_{i+1} - x_i})\omega x_i\right) = \sum_{i=1}^n \mathrm{Re}\left(\overline{(x_{i+1} - x_i)}\omega x_i\right)$, an inequality which we now reformulate in $\mathbb{C}^n$. Denote the $n \times n$-identity matrix by $\mathbf{I}$, and set

$$\mathbf{B} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \vdots \\ \vdots & & \ddots & \ddots & \\ & & & & 0 \\ 0 & & & & 1 \\ 1 & 0 & \cdots & & 0 \end{pmatrix} \in \mathbb{C}^{n\times n} \quad \text{and} \quad \mathbf{R} = \omega\mathbf{I} \in \mathbb{C}^{n\times n}. \tag{63}$$

Identifying $\mathbf{x} \in \mathbb{C}^n$ with $(x_1, \ldots, x_n) \in X^n$, we note that (61) means $0 \geq \mathrm{Re}\left(\left((\mathbf{B} - \mathbf{I})\mathbf{x}\right)^*\mathbf{R}\mathbf{x}\right)$; equivalently, $0 \geq \mathbf{x}^*(\mathbf{B}^* - \mathbf{I})\mathbf{R}\mathbf{x} + \mathbf{x}^*\mathbf{R}^*(\mathbf{B} - \mathbf{I})\mathbf{x}$. Set

$$\begin{aligned} \mathbf{C}_n &= (\mathbf{I} - \mathbf{B}^*)\mathbf{R} + \mathbf{R}^*(\mathbf{I} - \mathbf{B}) \\ &= \begin{pmatrix} (\omega + \overline{\omega}) & -\overline{\omega} & 0 & \cdots & 0 & -\omega \\ -\omega & (\omega + \overline{\omega}) & \ddots & & & 0 \\ 0 & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & & & & 0 \\ 0 & & & & (\omega + \overline{\omega}) & -\overline{\omega} \\ -\overline{\omega} & 0 & \cdots & 0 & -\omega & (\omega + \overline{\omega}) \end{pmatrix} \in \mathbb{C}^{n\times n}. \end{aligned} \tag{64}$$

Then

$$A_\theta \text{ is } n\text{-cyclically monotone} \Leftrightarrow \mathbf{C}_n \text{ is positive semidefinite on } \mathbb{C}^n. \tag{65}$$

Note that the matrix $\mathbf{C}_n$ is a circulant Toeplitz matrix; e.g., by [26, Exercise 5.8.12], the set of eigenvalues of $\mathbf{C}_n$ is

$$\Lambda_n = \left\{p(1), p(\zeta), \ldots, p(\zeta^{n-1})\right\}, \quad \text{where} \quad p\colon t \mapsto (\omega + \overline{\omega}) - \omega t - \overline{\omega}t^{n-1}, \tag{66}$$

and where ζ is an arbitrary nth root of unity. It will be convenient to work with

$$\zeta_n = \exp(-2\pi\mathrm{i}/n). \tag{67}$$

Then

$$\begin{aligned} \left(\forall k \in \{0, 1, \ldots, n-1\}\right) \quad p(\zeta_n^k) &= \omega + \overline{\omega} - \omega\zeta_n^k - \overline{\omega}(\zeta_n^k)^{n-1} \\ &= \omega + \overline{\omega} - \omega\zeta_n^k - \overline{\omega}(\zeta_n^{n-1})^k \\ &= \omega + \overline{\omega} - \omega\zeta_n^k - \overline{\omega}(\overline{\zeta_n})^k \\ &= \omega + \overline{\omega} - (\omega\zeta_n^k + \overline{\omega\zeta_n^k}) \\ &= 2\big(\cos(\theta) - \cos(2k\pi/n - \theta)\big). \end{aligned} \tag{68}$$

"$\Leftarrow$": Assume that $\theta \in [0, \pi/n]$. If $k \in \{1, 2, \ldots, n-1\}$, then $\theta \leq 2k\pi/n - \theta < 2\pi - \theta$ and (68) implies that $p(\zeta_n^k) \geq 0$. On the other hand, $p(1) = 0$. Altogether, every eigenvalue in Λ_n is nonnegative and the Hermitian matrix $\mathbf{C}_n$ is thus positive semidefinite. Therefore, by (65), A_θ is n-cyclically monotone.

"$\Rightarrow$": Assume that $\theta \in]\pi/(n+1), \pi/n]$. It suffices to show that A_θ is not $(n+1)$-cyclically monotone. Now (68) implies that $p(\zeta_{n+1}) = 2\big(\cos(\theta) - \cos(2\pi/(n+1) - \theta)\big) < 0$ since $0 < 2\pi/(n+1) - \theta < \theta$. In view of (66) and (65), we deduce that Λ_{n+1} contains a strictly negative eigenvalue, i.e., the matrix $\mathbf{C}_{n+1}$ is not positive semidefinite, and therefore A_θ is not $(n+1)$-cyclically monotone. ∎

Remark 7.2. The symmetric part of every continuous linear monotone operator is a subdifferential and hence cyclically monotone. Hence, higher-order n-cyclic monotonicity properties are not captured in the symmetric part. In other words, the analogue of Proposition 1.2 for n-cyclically monotone operators, where $n \in \{3, 4, \ldots\}$, is false: $A_{\pi/2}$ is not 3-cyclically monotone (by Proposition 7.1), yet its symmetric part $\big(A_{\pi/2}\big)_+ = 0$ is cyclically monotone.

PROPOSITION 7.3. *Let $n \in \{2, 3, \ldots\}$ and suppose that $\theta \in]\pi/(n+1), \pi/n]$. Then $F_{A_\theta, n+1} \equiv +\infty$.*

Proof. We shall utilize the following result on tridiagonal Toeplitz matrices; see [26, Example 7.2.5]:

> If $\alpha \in \mathbb{C} \smallsetminus \{0\}$, $\beta \in \mathbb{C}$, and $\gamma \in \mathbb{C} \smallsetminus \{0\}$, then the eigenvalues and the eigenvectors of the $n \times n$ matrix
>
> $$\begin{pmatrix} \beta & \alpha & 0 & \cdots & 0 \\ \gamma & \beta & \alpha & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \gamma & \beta & \alpha \\ 0 & \cdots & 0 & \gamma & \beta \end{pmatrix} \tag{69}$$
>
> are given by
>
> $$\lambda_k = \beta + 2\alpha\rho\cos\big(k\pi/(n+1)\big) \text{ and } \mathbf{x}_k = \begin{pmatrix} \rho\sin\big(k\pi/(n+1)\big) \\ \rho^2\sin\big(2k\pi/(n+1)\big) \\ \rho^3\sin\big(3k\pi/(n+1)\big) \\ \vdots \\ \rho^n\sin\big(nk\pi/(n+1)\big) \end{pmatrix}, \tag{70}$$
>
> respectively, where
>
> $$k \in \{1, 2, \ldots, n\} \quad \text{and} \quad \rho = \sqrt{\gamma/\alpha}. \tag{71}$$

We identify $\mathbb{R}^2$ with $\mathbb{C}$ as in the proof of Proposition 7.1, where we set $\omega = \exp(\mathrm{i}\theta)$. By (50), for an arbitrary $(x, u) \in \mathbb{R}^2 \times \mathbb{R}^2$, we have

$$\begin{aligned} F_{A_\theta, n+1}(x, u) &= \sup_{a_1, \ldots, a_n} \Big(\sum_{i=1}^{n-1} \langle a_{i+1} - a_i, A_\theta a_i \rangle \Big) + \langle x - a_n, A_\theta a_n \rangle + \langle a_1 - x, u \rangle + \langle x, u \rangle \\ &= \sup_{a_1, \ldots, a_n} \operatorname{Re}\Big(\Big(\sum_{i=1}^{n-1} \overline{(a_{i+1} - a_i)}\omega a_i \Big) + \overline{(-a_n)}\omega a_n + \overline{x}\omega a_n + \overline{a_1}u \Big) \qquad (72) \\ &= \sup_{\mathbf{a} \in \mathbb{C}^n} \tfrac{1}{2}\big(\mathbf{a}^*\mathbf{H}\mathbf{a} + (\overline{x}\omega a_n + x\overline{\omega a_n}) + (\overline{a_1}u + a_1\overline{u})\big), \end{aligned}$$

where $\mathbf{a} = (a_1, \ldots, a_n)^T \in \mathbb{C}^n$ and

$$\mathbf{H} = \begin{pmatrix} -(\omega+\overline{\omega}) & \overline{\omega} & 0 & \cdots & 0 \\ \omega & -(\omega+\overline{\omega}) & \overline{\omega} & 0 & \vdots \\ 0 & \ddots & \ddots & \ddots & \\ \vdots & & \omega & -(\omega+\overline{\omega}) & \overline{\omega} \\ 0 & \cdots & 0 & \omega & -(\omega+\overline{\omega}) \end{pmatrix} \in \mathbb{C}^{n\times n}. \tag{73}$$

By (70), the n eigenvalues of the Hermitian matrix $\mathbf{H}$ are given by

$$\begin{aligned} \big(\forall k \in \{1,\ldots,n\}\big) \quad \lambda_k &= -(\omega+\overline{\omega}) + 2\,\overline{\omega}\sqrt{\omega/\overline{\omega}}\cos\big(k\pi/(n+1)\big) \\ &= 2\big(\cos\big(k\pi/(n+1)\big) - \cos(\theta)\big). \end{aligned} \tag{74}$$

Since $0 < \pi/(n+1) < \theta \leq \pi/2$, we deduce that

$$\lambda_1 = 2\big(\cos(\pi/(n+1)) - \cos(\theta)\big) > 0. \tag{75}$$

Furthermore, since $\mathbf{H}$ is Hermitian, it can be unitarily diagonalized. There exists a unitary matrix $\mathbf{U} \in \mathbb{C}^{n\times n}$ such that $\mathbf{U}^*\mathbf{H}\mathbf{U} = \mathbf{D}$ is a diagonal matrix, with eigenvalues $\lambda_1, \ldots, \lambda_n$ on its diagonal. On one hand, changing variables via $\mathbf{a} = \mathbf{U}\mathbf{y}$, where $\mathbf{y} = (y_1, \ldots, y_n)^T \in \mathbb{C}^n$, we have

$$\mathbf{a}^*\mathbf{H}\mathbf{a} = \lambda_1|y_1|^2 + \cdots + \lambda_n|y_n|^2. \tag{76}$$

Note that if $\mathbf{y} = \tau(1, 0, \ldots, 0)^T$, then $\mathbf{a}^*\mathbf{H}\mathbf{a} = \lambda_1\tau^2$ is a convex quadratic in τ. On the other hand,

$$(\overline{x}\omega a_n + x\overline{\omega a_n}) + (\overline{a_1}x^* + a_1\overline{x^*}) \tag{77}$$

is $\mathbb{R}$-linear in $\mathbf{a}$, in $\mathbf{y}$, and in τ. Altogether, the supremum in (72) is equal to $+\infty$. This completes the proof. □

PROPOSITION 7.4. *Let $n \in \{2, 3, \ldots\}$ and suppose that $\theta = \pi/n$. Then $F_{A_\theta,n} = \iota_{\operatorname{gra} A_\theta} + \langle\cdot,\cdot\rangle$.*

Proof. Fix $(x, u) \in X \times X$. If $u = A_\theta x$, then $F_{A_\theta,n}(x, u) = \langle x, u\rangle$ by Fact 6.3. Thus assume that $u \neq A_\theta x$. Arguing as in the proof of Proposition 7.3, we see that

$$F_{A_\theta,n}(x,u) = \sup_{\mathbf{a}\in\mathbb{C}^{n-1}} \tfrac{1}{2}\big(\mathbf{a}^*\mathbf{H}\mathbf{a} + (\overline{x}\omega a_{n-1} + x\overline{\omega a_{n-1}}) + (\overline{a_1}u + a_1\overline{u})\big), \tag{78}$$

where $\omega = \exp(\mathrm{i}\theta) = \exp(\pi\mathrm{i}/n)$, $\mathbf{a} = (a_1, \ldots, a_{n-1})^T \in \mathbb{C}^{n-1}$, and

$$\mathbf{H} = \begin{pmatrix} -(\omega+\overline{\omega}) & \overline{\omega} & 0 & \cdots & 0 \\ \omega & -(\omega+\overline{\omega}) & \overline{\omega} & 0 & \vdots \\ 0 & \ddots & \ddots & \ddots & \\ \vdots & & \omega & -(\omega+\overline{\omega}) & \overline{\omega} \\ 0 & \cdots & 0 & \omega & -(\omega+\overline{\omega}) \end{pmatrix} \in \mathbb{C}^{(n-1)\times(n-1)}, \tag{79}$$

and where the eigenvalues $\mu_1, \ldots, \mu_{n-1}$ are given by (this is the counterpart of (74))

$$\big(\forall k \in \{1,\ldots,n-1\}\big) \quad \mu_k = 2\big(\cos(k\pi/n) - \cos(\theta)\big) \leq 0. \tag{80}$$

Note that $\mu_1 = 0$ and that, by (70),

$$\mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_{n-1} \end{pmatrix} = \begin{pmatrix} \exp(\pi\mathrm{i}/n)\sin(\pi/n) \\ \exp(2\pi\mathrm{i}/n)\sin(2\pi/n) \\ \vdots \\ \exp((n-1)\pi\mathrm{i}/n)\sin((n-1)\pi/n) \end{pmatrix} \tag{81}$$

is a corresponding eigenvector. Then $\mathbf{H}\mathbf{b} = \mathbf{0} \in \mathbb{C}^{n-1}$. Using $z\mathbf{b}$, where $z \in \mathbb{C}$, rather than the general vector $\mathbf{a}$ in (78), we estimate

(82)

$$\begin{aligned} &F_{A_\theta,n}(x,u) \\ &\geq \sup_{z\in\mathbb{C}} \operatorname{Re}\left(\overline{x}\omega z b_{n-1} + \overline{zb_1}u\right) \\ &= \sup_{z\in\mathbb{C}} \operatorname{Re}\left(\overline{x}\exp(\pi\mathrm{i}/n) z \exp((n-1)\pi\mathrm{i}/n)\sin((n-1)\pi/n) + \overline{z}\exp(-\pi\mathrm{i}/n)\sin(\pi/n)u\right) \\ &= \sin(\pi/n)\sup_{z\in\mathbb{C}} \operatorname{Re}\left(\overline{z}(u\exp(-\pi\mathrm{i}/n) - x)\right). \end{aligned}$$

Because $u \neq A_\theta x$, i.e., $u \neq \exp(\pi\mathrm{i}/n)x$ viewed in $\mathbb{C}$, we see that $u\exp(-\pi\mathrm{i}/n) - x \neq 0$. Thus, the last supremum is equal to $+\infty$. ∎

The following example will be utilized in Proposition 7.6.

Example 7.5. Suppose that $\theta \in [0, \pi/2[$. Then

$$F_{A_\theta,2}\colon \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}\colon (x,u) \mapsto \frac{1}{4\cos\theta}\|u + A_\theta^* x\|^2. \tag{83}$$

Proof. The symmetric part of A_θ is equal to $\cos(\theta)\,\mathrm{Id}$ and hence invertible. The result follows by combining Theorem 2.3(i) and Fact 2.2(vi). ∎

PROPOSITION 7.6. *Let $n \in \{2,3,\dots\}$ and suppose that $\theta \in]0, \pi/n[$. Then*

(84)

$$F_{A_\theta,n}\colon \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$$

$$(x,u) \mapsto \frac{\sin(n-1)\theta}{2\sin n\theta}\left(\|x\|^2 + \|u\|^2\right) + \frac{\sin\theta}{\sin n\theta}\langle x, A_\theta^{n-1}u\rangle \tag{85}$$

$$= \frac{\sin\theta}{2\sin n\theta}\left(\left(\frac{\sin(n-1)\theta}{\sin\theta} - 1\right)\left(\|x\|^2 + \|A_\theta^{n-1}u\|^2\right) + \|x + A_\theta^{n-1}u\|^2\right). \tag{86}$$

Proof. Observe that (86) is a direct consequence of (85). It suffices to verify (84)–(85), and we do this by induction on n. Fix $(x,u) \in \mathbb{R}^2 \times \mathbb{R}^2$. Consider the case when $n = 2$. Using Example 7.5 and the trigonometric identity $(\sin\theta)/(\sin 2\theta) = 1/(2\cos\theta)$, we obtain

$$F_{A_\theta,2}(x,u) = \frac{1}{4\cos\theta}\|u + A_\theta^* x\|^2 = \frac{\sin\theta}{2\sin 2\theta}\left(\|x\|^2 + \|u\|^2 + 2\langle u, A_\theta^* x\rangle\right), \tag{87}$$

which yields (85). We now assume that (85) holds for some $n \in \{2,3,\dots\}$, and we shall show that it also holds for $n+1$, provided that $\theta \in]0, \pi/(n+1)[$. Utilizing

Theorem 6.5 and trigonometric identities, we obtain

(88)
$$\begin{aligned}
&F_{A_\theta,n+1}(x,u)\\
&=\sup_{y\in X} F_{A_\theta,n}(y,u)+\langle x-y, A_\theta y\rangle\\
&=\sup_{y\in X} \frac{\sin(n-1)\theta}{2\sin n\theta}\big(\|y\|^2+\|u\|^2\big)+\frac{\sin\theta}{\sin n\theta}\langle y, A_\theta^{n-1}u\rangle+\langle A_\theta^* x, y\rangle-\langle y, A_\theta y\rangle\\
&=\sup_{y\in X} \Big(\frac{\sin(n-1)\theta}{2\sin n\theta}-\cos\theta\Big)\|y\|^2+\frac{\sin(n-1)\theta}{2\sin n\theta}\|u\|^2+\frac{\sin\theta}{\sin n\theta}\langle y, A_\theta^{n-1}u\rangle+\langle A_\theta^* x, y\rangle
\end{aligned}$$

(89)
$$=\sup_{y\in X} \frac{-\sin(n+1)\theta}{2\sin n\theta}\|y\|^2+\frac{\sin(n-1)\theta}{2\sin n\theta}\|u\|^2+\frac{\sin\theta}{\sin n\theta}\langle y, A_\theta^{n-1}u\rangle+\langle A_\theta^* x, y\rangle.$$

Since $\theta\in\left]0,\pi/(n+1)\right[$, the coefficient of $\|y\|^2$ is strictly negative, which shows that the quadratic function of y of which we take the supremum in (89) is strictly concave. Setting the derivative of this quadratic function equal to 0, we find that the unique maximizer in (89) is

$$\frac{\sin n\theta}{\sin(n+1)\theta}\Big(\frac{\sin\theta}{\sin n\theta}A_\theta^{n-1}u+A_\theta^* x\Big). \tag{90}$$

Combining this with (88) and (89), followed by simplification and utilization of trigonometric identities, we deduce that

$$F_{A_\theta,n+1}(x,u)=\frac{\sin n\theta}{2\sin(n+1)\theta}\big(\|x\|^2+\|u\|^2\big)+\frac{\sin\theta}{\sin(n+1)\theta}\langle x, A_\theta^n u\rangle, \tag{91}$$

and this completes the proof. □

Remark 7.7. Consider the setting of Proposition 7.6. Since $n\in\{2,3,\dots\}$ and since $\theta\in\left]0,\pi/n\right[$, we have $\theta\leq(n-1)\theta<\pi-\theta$ and thus $\sin(n-1)\theta\geq\sin\theta$. While it is clear from the definition that $F_{A_\theta,n}$ is convex (see (50)), we see this also directly from (86).

We have obtained complete knowledge of all Fitzpatrick functions. Let us summarize our findings.

THEOREM 7.8. *Let $\theta\in[0,\pi/2]$ and let A_θ be the rotator by θ in the Euclidean plane, i.e.,*

$$A_\theta=\begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix}. \tag{92}$$

(i) Case $\theta=0$. *Then $A_\theta=\mathrm{Id}=\nabla\frac{1}{2}\|\cdot\|^2$ is cyclically monotone, $F_{\mathrm{Id},\infty}=\frac{1}{2}\|\cdot\|^2\oplus\frac{1}{2}\|\cdot\|^2$, and*
(93)
$$\big(\forall n\in\{2,3,\dots\}\big)\quad F_{\mathrm{Id},n}\colon\mathbb{R}^2\times\mathbb{R}^2\to\mathbb{R}\colon(x,u)\mapsto\frac{n-1}{2n}\big(\|x\|^2+\|u\|^2\big)+\frac{1}{n}\langle x,u\rangle.$$

(ii) Case $\theta\in\left]0,\pi/2\right]$. *If $n\in\{2,3,\dots\}\cap[2,\pi/\theta[$, then A_θ is n-cyclically monotone and*
(94)
$$F_{A_\theta,n}\colon\mathbb{R}^2\times\mathbb{R}^2\to\mathbb{R}\colon(x,u)\mapsto\frac{\sin(n-1)\theta}{2\sin n\theta}\big(\|x\|^2+\|u\|^2\big)+\frac{\sin\theta}{\sin n\theta}\langle x, A_\theta^{n-1}u\rangle.$$

If π/θ is an integer, then A_θ is (π/θ)-cyclically monotone and

$$F_{A_\theta,\pi/\theta} = \iota_{\operatorname{gra} A_\theta} + \langle \cdot, \cdot \rangle. \tag{95}$$

If $n \in \{2, 3, \dots\} \cap]\pi/\theta, +\infty[$, then A_θ is not n-cyclically monotone and

$$F_{A_\theta,n} \equiv +\infty. \tag{96}$$

Proof. (i) This follows from Example 6.4 with $A = \mathrm{Id}$. (ii) is a direct consequence of Propositions 7.1, 7.3, 7.4, and 7.6. □

Remark 7.9. Theorem 7.8 greatly expands the knowledge about rotators and their Fitzpatrick functions. In previous work [2], only rotators by 0 or by π/n, where $n \in \{2, 3, \dots\}$, were considered. In that restricted setting, item (i) of Theorem 7.8 was known [2, Example 4.4]. It was also known that $A_{\pi/n}$ is n-cyclically monotone but not $(n+1)$-cyclically monotone [2, Example 4.6]. The formula (95) was known only for $\theta = \pi/2$ [2, Example 4.5], and formula (96) was known only for $n \in \{2, 3, 4\}$ [2, Remark 4.7].

Acknowledgment. The authors wish to thank Sedi Bartz, Simeon Reich, and two anonymous referees for their pertinent comments which led to considerable improvements over the original version of this manuscript.

REFERENCES

[1] E. Asplund, *A monotone convergence theorem for sequences of nonlinear mappings*, in Nonlinear Functional Analysis, Proceedings of Symposia in Pure Mathematics, Vol. 18, Part 1, AMS, Providence, RI, 1970, pp. 1–9.

[2] S. Bartz, H. H. Bauschke, J. M. Borwein, S. Reich, and X. Wang, *Fitzpatrick functions, cyclic monotonicity, and Rockafellar's antiderivative*, Nonlinear Anal., 66 (2007), pp. 1198–1223.

[3] H. H. Bauschke, *Projection Algorithms and Monotone Operators*, Ph.D. thesis, Department of Mathematics, Simon Fraser University, Burnaby, BC, Canada, 1996; available online at http://www.cecm.sfu.ca/preprints/1996pp.html.

[4] H. H. Bauschke, *The composition of finitely many projections onto closed convex sets in Hilbert space is asymptotically regular*, Proc. Amer. Math. Soc., 131 (2003), pp. 141–146.

[5] H. H. Bauschke and J. M. Borwein, *Maximal monotonicity of dense type, local maximal monotonicity, and monotonicity of the conjugate are all the same for continuous linear operators*, Pacific J. Math., 189 (1999), pp. 1–20.

[6] H. H. Bauschke, D. A. McLaren, and H. S. Sendov, *Fitzpatrick functions: Inequalities, examples and remarks on a problem by S. Fitzpatrick*, J. Convex Anal., 13 (2006), pp. 499–523.

[7] H. H. Bauschke and S. Simons, *Stronger maximal monotonicity properties of linear operators*, Bull. Austral. Math. Soc., 60 (1999), pp. 163–174.

[8] J. M. Borwein, *Maximal monotonicity via convex analysis*, J. Convex Anal., 13 (2006), pp. 561–586.

[9] J. M. Borwein and Q. J. Zhu, *Techniques of Variational Analysis*, Springer-Verlag, New York, 2005.

[10] H. Brézis, *Opérateurs Maximaux Monotones et Semi-Groupes de Contractions dans les Espaces de Hilbert*, North–Holland, Amsterdam, 1973.

[11] H. Brézis and A. Haraux, *Image d'une somme d'opérateurs monotones et applications*, Israel J. Math., 23 (1976), pp. 165–186.

[12] R. S. Burachik and S. Fitzpatrick, *On a family of convex functions associated to subdifferentials*, J. Nonlinear Convex Anal., 6 (2005), pp. 165–171.

[13] R. S. Burachik and S. Fitzpatrick, *Corrigendum to: "On a family of convex functions associated to subdifferentials,"* J. Nonlinear Convex Anal., 6 (2005), p. 535.

[14] R. S. Burachik and B. F. Svaiter, *Maximal monotone operators, convex functions and a special family of enlargements*, Set-Valued Anal., 10 (2002), pp. 297–316.

[15] R. S. Burachik and B. F. Svaiter, *Maximal monotonicity, conjugation and the duality product*, Proc. Amer. Math. Soc., 131 (2003), pp. 2379–2383.

[16] S. Cheng and Y. Tian, *Two sets of new characterizations for normal and EP matrices*, Linear Algebra Appl., 375 (2003), pp. 181–195.
[17] S. Fitzpatrick, *Representing monotone operators by convex functions*, in Workshop/Miniconference on Functional Analysis and Optimization (Canberra, 1988), Proc. Centre Math. Anal. Austral. Nat. Univ. 20, Australian National University, Canberra, Australia, 1988, pp. 59–65.
[18] Y. García, M. Lassonde, and J. P. Revalski, *Extended sums and extended compositions of monotone operators*, J. Convex Anal., 13 (2006), pp. 721–738.
[19] K. Goebel and S. Reich, *Uniform Convexity, Hyperbolic Geometry, and Nonexpansive Mappings*, Marcel Dekker, New York, 1984.
[20] C. W. Groetsch, *Generalized Inverses of Linear Operators*, Marcel Dekker, New York, 1977.
[21] P. R. Halmos, *A Hilbert Space Problem Book*, Van Nostrand, Princeton, NJ, 1967.
[22] A. N. Iusem, *On some properties of paramonotone operators*, J. Convex Anal., 5 (1998), pp. 269–278.
[23] E. Kreyszig, *Introductory Functional Analysis with Applications*, John Wiley & Sons, New York, 1989.
[24] J.-E. Martínez-Legaz and B. F. Svaiter, *Monotone operators representable by l.s.c. convex functions*, Set-Valued Anal., 13 (2005), pp. 21–46.
[25] J.-E. Martínez-Legaz and M. Théra, *A convex representation of maximal monotone operators*, J. Nonlinear Convex Anal., 2 (2001), pp. 243–247.
[26] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra*, SIAM, Philadelphia, 2000.
[27] B. S. Mordukhovich, *Variational Analysis and Generalized Differentiation* I, Springer-Verlag, Berlin, 2006.
[28] B. S. Mordukhovich and Y. Shao, *Nonsmooth sequential analysis in Asplund spaces*, Trans. Amer. Math. Soc., 348 (1996), pp. 1235–1280.
[29] J. P. Penot, *The relevance of convex analysis for the study of monotonicity*, Nonlinear Anal., 58 (2004), pp. 855–871.
[30] R. R. Phelps and S. Simons, *Unbounded linear monotone operators on nonreflexive Banach spaces*, J. Convex Anal., 5 (1998), pp. 303–328.
[31] S. Reich and S. Simons, *Fenchel duality, Fitzpatrick functions and the Kirszbraun-Valentine extension theorem*, Proc. Amer. Math. Soc., 133 (2005), pp. 2657–2660.
[32] R. T. Rockafellar, *On the maximal monotonicity of subdifferential mappings*, Pacific J. Math., 33 (1970), pp. 209–216.
[33] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[34] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.
[35] R. T. Rockafellar and R. J. B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
[36] S. Simons, *Minimax and Monotonicity*, Lecture Notes in Math. 1693, Springer-Verlag, Berlin, 1998.
[37] S. Simons, *Dualized and scaled Fitzpatrick functions*, Proc. Amer. Math. Soc., 134 (2006), pp. 2983–2987.
[38] S. Simons, *Positive sets and monotone sets*, J. Convex Anal., 14 (2007), pp. 297–318.
[39] S. Simons, *LC-functions and maximal monotonicity*, J. Nonlinear Convex Anal., 7 (2006), pp. 123–137.
[40] S. Simons, *The Fitzpatrick function and nonreflexive spaces*, J. Convex Anal., 13 (2006), pp. 861–881.
[41] S. Simons and C. Zălinescu, *A new proof for Rockafellar's characterization of maximal monotone operators*, Proc. Amer. Math. Soc., 132 (2004), pp. 2969–2972.
[42] S. Simons and C. Zălinescu, *Fenchel duality, Fitzpatrick functions and maximal monotonicity*, J. Nonlinear Convex Anal., 6 (2005), pp. 1–22.
[43] C. Zălinescu, *Convex Analysis in General Vector Spaces*, World Scientific, River Edge, NJ, 2002.

SIAM J. OPTIM.
Vol. 18, No. 3, pp. 810–833

DIRECTIONAL REGULARITY AND METRIC REGULARITY*

ARAM V. ARUTYUNOV†, EVGENIY R. AVAKOV‡, AND ALEXEY F. IZMAILOV§

Abstract. For general constraint systems in Banach spaces, we present the directional stability theorem based on the appropriate generalization of the directional regularity condition, suggested earlier in [A. V. Arutyunov and A. F. Izmailov, *Math. Oper. Res.*, 31 (2006), pp. 526–543]. This theorem contains Robinson's stability theorem but does not reduce to it. Furthermore, we develop the related concept of directional metric regularity which is stable subject to small Lipschitzian perturbations of the constraint mapping, and which is equivalent to directional regularity for sufficiently smooth mappings. Finally, we discuss some applications in sensitivity theory.

Key words. metric regularity, Robinson's constraint qualification, directional regularity, directional metric regularity, feasible arc, sensitivity

AMS subject classifications. 49K27, 49K40, 90C31

DOI. 10.1137/060651616

1. Introduction. Directional regularity. Let Σ be a topological space, X and Y be Banach spaces, and Q be a fixed closed set in Y. Consider a smooth mapping $F : \Sigma \times X \to Y$ (our smoothness hypotheses will be specified below), and set

$$D(\sigma) = \{x \in X \mid F(\sigma, x) \in Q\} \tag{1.1}$$

with $\sigma \in \Sigma$ playing the role of a parameter. For a given (base) parameter value $\sigma_0 \in \Sigma$, fix $x_0 \in D(\sigma_0)$. In this paper we are concerned with the following question: For which $(\sigma, x) \in \Sigma \times X$ close to (σ_0, x_0), and under which assumptions can $\operatorname{dist}(x, D(\sigma))$ be estimated from above via the "residual" of constraints in (1.1), that is, via $\operatorname{dist}(F(\sigma, x), Q)$? Here $\operatorname{dist}(z, S) = \inf_{s \in S} \|z - s\|$ stands for the distance from a point z to a set S.

The answer to this question is well known provided Q is convex and the so-called Robinson's constraint qualification (CQ) is satisfied at x_0 for the mapping $F(\sigma_0, \cdot)$, that is,

$$0 \in \operatorname{int}\left(F(\sigma_0, x_0) + \operatorname{im}\frac{\partial F}{\partial x}(\sigma_0, x_0) - Q\right), \tag{1.2}$$

where $\operatorname{int} S$ is the interior of a set S, and $\operatorname{im}\Lambda$ is the range (image space) of a linear operator Λ. According to Robinson's stability theorem [21] (see also [3, Theorem 2.87]),

*Received by the editors February 6, 2006; accepted for publication (in revised form) December 4, 2006; published electronically October 4, 2007. The research of the first two authors was supported by the Russian Foundation for Basic Research grants 05-01-00193 and 05-01-00275, and by RF President's grants NS-5344.2006.1 and NS-5813.2006.1 for the state support of leading scientific schools. The third author was supported by the Russian Foundation for Basic Research grant 04-01-00341, by RF President's grant NS-9394.2006.1 for the state support of leading scientific schools, and by RF President's grant MD-2723.2005.1 for the state support of young doctors of sciences.

http://www.siam.org/journals/siopt/18-3/65161.html

†Peoples' Friendship University, Miklukho-Maklaya Str. 6, 117198 Moscow, Russia (arutun@orc.ru).

‡Institute for Control Problems RAS, Profsoyuznaya Str. 65, 117806 Moscow, Russia (era@maxmin.ru).

§Faculty of Computational Mathematics and Cybernetics, Department of Operations Research, Moscow State University, Leninskiye Gori, GSP-2, 119992 Moscow, Russia (izmaf@ccas.ru).

under these assumptions there exists a constant $c > 0$ such that the estimate

$$\operatorname{dist}(x, D(\sigma)) \le c \operatorname{dist}(F(\sigma, x), Q) \tag{1.3}$$

holds *for all* $(\sigma, x) \in \Sigma \times X$ close enough to (σ_0, x_0).

In its turn, estimate (1.3) serves as a motivation for the very important concept of metric regularity. Apparently, the term "metric regularity" appeared for the first time in [4], but the concept dates back to earlier works [14, 20, 10] (or even to classical works [15, 12]; see also [6, 5, 19]), and it finds multiple applications in modern variational analysis. Specifically, the mapping $F : X \to Y$ is said to be metrically regular at $x_0 \in F^{-1}(Q)$ with respect to Q if there exists a constant $c > 0$ such that the estimate

$$\operatorname{dist}(x, F^{-1}(Q + y)) \le c \operatorname{dist}(F(x) - y, Q) \tag{1.4}$$

holds for all $(x, y) \in X \times Y$ close enough to $(x_0, 0)$. Note that (1.4) is nothing else but the estimate (1.3) for $F(\sigma, x) = F(x) - \sigma$ and $\sigma = y$, i.e., for the special parametrization of the mapping F in question (the "right-hand side" perturbations). Thus, by Robinson's stability theorem, if Q is convex, then Robinson's CQ

$$0 \in \operatorname{int}(F(x_0) + \operatorname{im} F'(x_0) - Q)$$

implies metric regularity of F at x_0 with respect to Q. Moreover, as is well known (see, e.g., [3, Proposition 2.89]), under the appropriate smoothness hypothesis, the converse implication is true as well, and thus, metric regularity and Robinson's CQ are actually equivalent.

For more recent developments and extensions of the metric regularity theory, see, e.g., [17, 22, 13, 18, 16] and references therein.

In particular, if Robinson's CQ does not hold, one cannot expect a smooth mapping to be metrically regular. Accordingly, for a parametric mapping F, estimate (1.3) for all (σ, x) close enough to (σ_0, x_0) cannot be guaranteed if (1.2) does not hold. However, we demonstrate below that under the regularity condition weaker than (1.2), estimate (1.3) is still valid but possibly not *for all* (σ, x) in a neighborhood of (σ_0, x_0); the set of appropriate (σ, x) will be specified. To this end, we give the following definition.

DEFINITION 1.1. *The mapping* $F(\sigma_0, \cdot) : X \to Y$ *is* regular at $x_0 \in D(\sigma_0)$ in a direction $\bar{y} \in Y$ *if*

$$0 \in \operatorname{int}\left(F(\sigma_0, x_0) + \operatorname{im} \frac{\partial F}{\partial x}(\sigma_0, x_0) - \operatorname{cone}\{\bar{y}\} - Q\right), \tag{1.5}$$

where $\operatorname{cone} S$ *stands for the conic hull of a set* S.

Note that for $\bar{y} = 0$, condition (1.5) reduces to Robinson's CQ (1.2). Moreover, if the latter is satisfied, the directional regularity condition (1.5) holds in any direction $\bar{y} \in Y$, including $\bar{y} = 0$.

Condition (1.5) and the corresponding directional stability result were first suggested in [1] for the case of finite-dimensional Y. However, the estimate obtained in [1, Theorem 4.1] is somewhat weaker than (1.3). This is a consequence of the general framework adopted in [1]. Specifically, the authors first consider the case of equality constraints and direct set constraints with a closed convex set P and prove the directional stability theorem with the estimate to the solution set only from points in P. Then they reduce (1.1) to this setting. On the other hand, the proof of the

directional stability theorem in [1] is very concise and clear, and, in particular, it does not appeal to any set-valued analysis. At the same time, the assumption $\dim Y < \infty$ cannot be dropped in that proof (and hence, in all the results obtained in [1]) because the argument there employs (the completely finite-dimensional) Brouwer's fixed point theorem. (We note, however, that in [1, Theorem 4.1], X can actually be just a normed linear space, not necessarily complete.)

In section 2, we prove the directional stability theorem (Theorem 2.3) under the same set of assumptions as in [1], but with the resulting estimate of the "proper" form (1.3), and for a (possibly infinite-dimensional) Banach space Y. In particular, Theorem 2.3 contains Robinson's stability theorem but does not reduce to it, in general.

Furthermore, in section 3, for a nonparametric mapping, we develop the directional metric regularity concept suggested by Theorem 2.3. In Theorem 3.2 we demonstrate that this property is stable subject to small Lipschitzian perturbations of F. This result combined with Theorem 2.3 implies the equivalence of directional regularity and directional metric regularity for sufficiently smooth mappings.

Finally, in section 4, we demonstrate that Theorem 2.3 can be used in order to directly obtain various stability results widely used in sensitivity analysis [3]. Specifically, assuming that Σ is a normed linear space, we consider the case when for a given direction $d \in \Sigma$ it holds that

$$0 \in \operatorname{int}\left(F(\sigma_0,\, x_0) + \operatorname{im}\frac{\partial F}{\partial x}(\sigma_0,\, x_0) + \operatorname{cone}\left\{\frac{\partial F}{\partial \sigma}(\sigma_0,\, x_0)d\right\} - Q\right). \tag{1.6}$$

Note that (1.6) is a particular case of (1.5) for $\bar{y} = -\frac{\partial F}{\partial \sigma}(\sigma_0,\, x_0)d$. On the other hand, (1.5) can be interpreted as (1.6) with F replaced by $\widetilde{F}(\sigma,\, y,\, x) = F(\sigma,\, x) - y$, where $y \in Y$ is regarded as an additional parameter, and with $d = (0,\, \bar{y}) \in \Sigma \times Y$.

In the context of mathematical programming problems, (1.6) is known as Gollan's condition [11]. It was extended to the general case in [2] (see also [3, Theorem 4.9]). Moreover, in parametric optimization, *this* condition (which is a particular case of (1.5)) is commonly known as the directional regularity condition. Taking into account the relations between the two conditions discussed above, the authors prefer to use the same name for the property stated in Definition 1.1. Note, however, that unlike (1.6), (1.5) does not depend on a specific parametrization at all: it is entirely a property of the unperturbed constraints. This makes our directional regularity particularly useful for unification of some diverse developments, like those based on Robinson's CQ and on customary directional regularity (1.6).

2. Directional stability theorem. In what follows, we shall need some equivalent formulations of the directional regularity condition introduced in Definition 1.1.

PROPOSITION 2.1. *Let Q be closed and convex.*

Then condition (1.5) *is equivalent to each of the following three conditions:*

$$\operatorname{cone}\{\bar{y}\} \cap \operatorname{int}\left(F(\sigma_0,\, x_0) + \operatorname{im}\frac{\partial F}{\partial x}(\sigma_0,\, x_0) - Q\right) \neq \emptyset, \tag{2.1}$$

$$\bar{y} \in \operatorname{int}\left(\operatorname{im}\frac{\partial F}{\partial x}(\sigma_0,\, x_0) - R_Q(F(\sigma_0,\, x_0))\right), \tag{2.2}$$

and

$$\operatorname{im}\frac{\partial F}{\partial x}(\sigma_0,\, x_0) - \operatorname{cone}\{\bar{y}\} - R_Q(F(\sigma_0,\, x_0)) = Y, \tag{2.3}$$

where $R_S(z) = \operatorname{cone}(S - z)$ *stands for the radial cone to a set* S *at a point* $z \in S$.

Note that condition (2.1) can be expressed in the following form: there exists $\theta \geq 0$ such that

$$\theta\bar{y} \in \operatorname{int}\left(F(\sigma_0,\, x_0) + \operatorname{im}\frac{\partial F}{\partial x}(\sigma_0,\, x_0) - Q\right). \tag{2.4}$$

Proof. (1.5) $\Rightarrow$ (2.1). The proof of this implication is almost identical to that of the corresponding assertion in [3, Theorem 4.9] (see the argument showing that (4.12) implies (4.13)). Define the multifunction $\Psi : X \times \mathbf{R} \to 2^Y$:

$$\Psi(x,\, \theta) = \begin{cases} F(\sigma_0,\, x_0) + \frac{\partial F}{\partial x}(\sigma_0,\, x_0)x - \theta\bar{y} - Q & \text{if } \theta \geq 0, \\ \emptyset & \text{if } \theta = 0. \end{cases} \tag{2.5}$$

Evidently, Ψ is a closed convex multifunction (that is, $\operatorname{graph}\Psi$ is a closed convex set; see, e.g., [3, p. 55]), and

$$\Psi(X \times \mathbf{R}) = F(\sigma_0,\, x_0) + \operatorname{im}\frac{\partial F}{\partial x}(\sigma_0,\, x_0) - \operatorname{cone}\{\bar{y}\} - Q,$$

and thus, (1.5) means that $0 \in \operatorname{int}\Psi(X \times \mathbf{R})$. Furthermore, $0 \in \Psi(0,\, 0)$, and hence, by the generalized open mapping theorem [3, Theorem 2.70] and by (2.5), $0 \in \operatorname{int}\Psi(X \times [0,\, 1])$. This means that there exists $\delta > 0$ such that

$$\begin{aligned} B_\delta(0) &\subset \Psi(X \times [0,\, 1]) \\ &= F(\sigma_0,\, x_0) + \operatorname{im}\frac{\partial F}{\partial x}(\sigma_0,\, x_0) - \{\theta\bar{y} \mid \theta \in [0,\, 1]\} - Q, \end{aligned} \tag{2.6}$$

where $B_\delta(z)$ stands for the ball centered at z and of radius δ.

Fix $\tilde{\delta} > 0$ small enough so that $\tilde{\delta}\bar{y} \in B_\delta(0)$. Then inclusion (2.6) implies that there exists $\tilde{\theta} \in [0,\, 1]$ such that

$$\tilde{\delta}\bar{y} \in F(\sigma_0,\, x_0) + \operatorname{im}\frac{\partial F}{\partial x}(\sigma_0,\, x_0) - \tilde{\theta}\bar{y} - Q,$$

and hence,

$$(\tilde{\delta} + \tilde{\theta})\bar{y} \in F(\sigma_0,\, x_0) + \operatorname{im}\frac{\partial F}{\partial x}(\sigma_0,\, x_0) - Q.$$

The set in the right-hand side of the latter inclusion is convex and contains 0, and thus

$$\{(\tilde{\delta} + \tilde{\theta})\theta\bar{y} \mid \theta \in [0,\, 1]\} \subset F(\sigma_0,\, x_0) + \operatorname{im}\frac{\partial F}{\partial x}(\sigma_0,\, x_0) - Q.$$

Then inclusion (2.6) implies that

$$\begin{aligned} B_\delta(0) &\subset F(\sigma_0,\, x_0) + \operatorname{im}\frac{\partial F}{\partial x}(\sigma_0,\, x_0) - \bar{y} + \{\theta\bar{y} \mid \theta \in [0,\, 1]\} - Q \\ &\subset (1 + 1/(\tilde{\delta} + \tilde{\theta}))\left(F(\sigma_0,\, x_0) + \operatorname{im}\frac{\partial F}{\partial x}(\sigma_0,\, x_0) - Q\right) - \bar{y}. \end{aligned}$$

It follows that

$$\begin{aligned} B_{\delta\theta}(\theta\bar{y}) &= \theta\bar{y} + B_{\delta\theta}(0) \\ &\subset F(\sigma_0,\, x_0) + \operatorname{im}\frac{\partial F}{\partial x}(\sigma_0,\, x_0) - Q \end{aligned}$$

holds with $\theta = (1 + 1/(\tilde{\delta} + \tilde{\theta}))^{-1} > 0$, and (2.4) (and hence (2.1)) is thus proved.

(2.1) $\Rightarrow$ (2.2). Since $Q - F(\sigma_0, x_0) \subset R_Q(F(\sigma_0, x_0))$, condition (2.1) clearly implies that

$$\operatorname{int}\left(\operatorname{im}\frac{\partial F}{\partial x}(\sigma_0, x_0) - R_Q(F(\sigma_0, x_0))\right) \neq \emptyset.$$

Suppose that (2.2) does not hold. Then by the first separation theorem [3, Theorem 2.13], there exists $\mu \in Y^*$ such that

$$\langle \mu, \bar{y}\rangle \leq \langle \mu, \eta\rangle \quad \forall\, \eta \in \operatorname{im}\frac{\partial F}{\partial x}(\sigma_0, x_0) - R_Q(F(\sigma_0, x_0)).$$

This evidently implies that

$$\langle \mu, \theta\bar{y}\rangle \leq 0 \leq \langle \mu, y\rangle \quad \forall\, \theta \geq 0,\ \forall\, y \in F(\sigma_0, x_0) + \operatorname{im}\frac{\partial F}{\partial x}(\sigma_0, x_0) - Q,$$

where the inclusion $Q - F(\sigma_0, x_0) \subset R_Q(F(\sigma_0, x_0))$ was again taken into account. Hence, μ separates $\operatorname{cone}\{\bar{y}\}$ and $F(\sigma_0, x_0) + \operatorname{im}\frac{\partial F}{\partial x}(\sigma_0, x_0) - Q$, and according to the first separation theorem [3, Theorem 2.13], this contradicts (2.1).

(2.2) $\Rightarrow$ (2.3). By (2.2), there exists $\delta > 0$ such that

$$\begin{aligned}\bar{y} + B_\delta(0) &= B_\delta(\bar{y})\\ &\subset \operatorname{im}\frac{\partial F}{\partial x}(\sigma_0, x_0) - R_Q(F(\sigma_0, x_0)),\end{aligned}$$

and hence

$$\begin{aligned}B_\delta(0) &\subset \operatorname{im}\frac{\partial F}{\partial x}(\sigma_0, x_0) - \bar{y} - R_Q(F(\sigma_0, x_0))\\ &\subset \operatorname{im}\frac{\partial F}{\partial x}(\sigma_0, x_0) - \operatorname{cone}\{\bar{y}\} - R_Q(F(\sigma_0, x_0)).\end{aligned}$$

Thus,

$$0 \in \operatorname{int}\left(\operatorname{im}\frac{\partial F}{\partial x}(\sigma_0, x_0) - \operatorname{cone}\{\bar{y}\} - R_Q(F(\sigma_0, x_0))\right)$$

holds, which evidently implies (2.3).

(2.3) $\Rightarrow$ (1.5). The proof of this implication is almost identical to that of the corresponding assertion in [3, Proposition 2.95] (see the argument showing that (2.180) implies (2.178)). Define the multifunction $\Psi : X \times \mathbf{R} \times \mathbf{R} \to 2^Y$:

$$\Psi(x, \theta, \tau) = \begin{cases} \frac{\partial F}{\partial x}(\sigma_0, x_0)x - \theta\bar{y} - \tau(Q - F(\sigma_0, x_0)) & \text{if } \theta \geq 0,\ \tau \geq 0,\\ \emptyset & \text{otherwise.}\end{cases} \tag{2.7}$$

Evidently, Ψ is a closed convex multifunction, and

$$\Psi(X \times \mathbf{R} \times \mathbf{R}) = \operatorname{im}\frac{\partial F}{\partial x}(\sigma_0, x_0) - \operatorname{cone}\{\bar{y}\} - R_Q(F(\sigma_0, x_0)),$$

and thus, (2.3) implies that $0 \in \operatorname{int}\Psi(X \times \mathbf{R} \times \mathbf{R})$. Furthermore, $0 \in \Psi(0, 0, 0)$, and hence, by the generalized open mapping theorem [3, Theorem 2.70] and by (2.7),

$$0 \in \operatorname{int}\Psi(X \times \mathbf{R}_+ \times [0, 1]). \tag{2.8}$$

On the other hand,

$$\begin{aligned}\Psi(X \times \mathbf{R}_+ \times [0,\,1]) &= \operatorname{im} \frac{\partial F}{\partial x}(\sigma_0,\,x_0) - \operatorname{cone}\{\bar y\} \\ &\quad -\{\tau(q - F(\sigma_0,\,x_0)) \mid \tau \in [0,\,1],\, q \in Q\} \\ &= F(\sigma_0,\,x_0) + \operatorname{im} \frac{\partial F}{\partial x}(\sigma_0,\,x_0) - \operatorname{cone}\{\bar y\} \\ &\quad -\{\tau q + (1-\tau)F(\sigma_0,\,x_0) \mid \tau \in [0,\,1],\, q \in Q\} \\ &\subset F(\sigma_0,\,x_0) + \operatorname{im} \frac{\partial F}{\partial x}(\sigma_0,\,x_0) - \operatorname{cone}\{\bar y\} - Q,\end{aligned}$$

where the convexity of Q was taken into account. It follows that (2.8) implies (1.5). □

We shall also need the following proposition.

PROPOSITION 2.2. *Let Q be convex, $y_0 \in Q$.*

Then for any $\bar y \in Y$ and any $\delta_1 > 0$, $\delta_2 > 0$ there exist $\varepsilon > 0$ and $\delta > 0$ such that

$$(Q - \operatorname{cone} B_\delta(\bar y)) \cap B_\varepsilon(y_0) \subset Q \cap B_{\delta_1}(y_0) - \operatorname{cone} B_{\delta_2}(\bar y). \tag{2.9}$$

Proof. First suppose that $\bar y \in T_Q(y_0)$, where $T_Q(y_0) = \operatorname{cl} R_Q(y_0)$ is the tangent cone to Q at y_0. We claim that in this case

$$y_0 \in \operatorname{int}(Q \cap B_{\delta_1}(y_0) - \operatorname{cone} B_{\delta_2}(\bar y)), \tag{2.10}$$

and hence, (2.9) evidently holds with an arbitrary $\delta > 0$ and a sufficiently small $\varepsilon > 0$.

Indeed, the interior of the set in the right-hand side of (2.10) is nonempty, and if (2.10) does not hold then by the first separation theorem [3, Theorem 2.13] there exists $\mu \in Y^*$ such that

$$\langle \mu,\, y\rangle \geq \langle \mu,\, y_0\rangle \quad \forall\, y \in Q \cap B_{\delta_1}(y_0) - \operatorname{cone} B_{\delta_2}(\bar y). \tag{2.11}$$

Then evidently

$$\langle \mu,\, \eta\rangle \geq 0 \quad \forall\, \eta \in T_Q(y_0). \tag{2.12}$$

On the other hand, for any $y \in B_{\delta_2}(0)$ such that $\langle \mu,\, y\rangle > 0$, from (2.11) we obtain

$$\begin{aligned}\langle \mu,\, y_0 - \bar y\rangle &> \langle \mu,\, y_0 - (\bar y + y)\rangle \\ &\geq \langle \mu,\, y_0\rangle,\end{aligned}$$

and thus $\langle \mu,\, \bar y\rangle < 0$ which contradicts (2.12) (recall that $\bar y \in T_Q(y_0)$).

Now let $\bar y \notin T_Q(y_0)$. Since $T_Q(y_0)$ is closed, by the second separation theorem [3, Theorem 2.14] we then obtain the existence of $\mu \in Y^*$ such that (2.12) holds and $\langle \mu,\, \bar y\rangle < 0$.

Consider arbitrary sequences $\{q^k\} \subset Q$, $\{\eta^k\} \subset Y$ and a sequence of real numbers $\{t_k\}$ such that $t_k \geq 0$ for all k and $\{q^k - t_k\eta^k\} \to y_0$. Hence

$$\langle \mu,\, q^k - y_0\rangle + t_k(-\langle \mu,\, \eta^k\rangle) = \langle \mu,\, q^k - t_k\eta^k - y_0\rangle \to 0.$$

Note that $q^k - y_0 \in R_Q(y_0) \subset T_Q(y_0)$, and (2.12) implies that the first term in the left-hand side is nonnegative for all k. Furthermore, inequality $\langle \mu,\, \bar y\rangle < 0$ implies that the second term in the left-hand side is nonnegative as well for all k large enough,

and hence, $t_k \to 0$. The latter implies that $\{q^k\} \to y_0$. Thus, $q^k - t_k\eta^k \in Q \cap B_{\delta_1}(y_0) - \operatorname{cone} B_{\delta_2}(\bar{y})$ for all k large enough. This proves the needed inclusion (2.9) with sufficiently small $\varepsilon > 0$ and $\delta > 0$. □

We are now ready to prove the main result of this section.

THEOREM 2.3. *Let Q be closed and convex, and let $x_0 \in D(\sigma_0)$. Let F be continuous at (σ_0, x_0) and Fréchet-differentiable with respect to x near (σ_0, x_0), and let its derivative with respect to x be continuous at (σ_0, x_0).*

If the mapping $F(\sigma_0, \cdot)$ is regular at x_0 in a direction $\bar{y} \in Y$, then there exist a neighborhood U of σ_0 and $\varepsilon > 0$, $\delta > 0$, and $c > 0$ such that the estimate (1.3) *holds for all $(\sigma, x) \in U \times B_\varepsilon(x_0)$ satisfying the inclusion*

$$F(\sigma, x) \in Q - \operatorname{cone} B_\delta(\bar{y}). \tag{2.13}$$

Proof. From the equivalent form (2.1) of the directional regularity condition it evidently follows that there exists $\bar{\eta} \in \operatorname{cone}\{\bar{y}\}$ such that

$$\bar{\eta} \in \operatorname{int}\left(F(\sigma_0, x_0) + \operatorname{im}\frac{\partial F}{\partial x}(\sigma_0, x_0) - Q\right). \tag{2.14}$$

Note that if $\bar{y} = 0$ then necessarily $\bar{\eta} = 0$.

Define the multifunction $\bar{\mathcal{F}} : X \to 2^Y$:

$$\bar{\mathcal{F}}(\xi) = F(\sigma_0, x_0) + \frac{\partial F}{\partial x}(\sigma_0, x_0)\xi - Q.$$

According to (2.14), there exists $\bar{\xi} \in X$ such that $\bar{\mathcal{F}}(\bar{\xi}) = \bar{\eta}$, and moreover, by the Robinson–Ursescu stability theorem [23, 20] (see also [3, Theorem 2.83]) it follows that the multifunction $\bar{\mathcal{F}}$ is metrically regular at $(\bar{\xi}, \bar{\eta})$.

Fix $\bar{\varepsilon} > 0$. For each mapping $G : X \to Y$, define the multifunction $\mathcal{F}_G : X \to 2^Y$,

$$\mathcal{F}_G(\xi) = F(\sigma_0, x_0) + G(\xi) - Q.$$

Note that $\mathcal{F}_{\frac{\partial F}{\partial x}(\sigma_0, x_0)} = \bar{\mathcal{F}}$, and hence, by [3, Theorem 2.84] it follows that there exist $\bar{l} > 0$, $\delta > 0$, and $\bar{c} > 0$ such that the estimate

$$\begin{aligned}\operatorname{dist}(\bar{\xi}, \mathcal{F}_G^{-1}(y)) \le \bar{c}\operatorname{dist}(G(\bar{\xi}) - y, Q - F(\sigma_0, x_0))\\ \forall\, y \in B_\delta\left(\bar{\eta} - \frac{\partial F}{\partial x}(\sigma_0, x_0)\bar{\xi} + G(\bar{\xi})\right)\end{aligned} \tag{2.15}$$

holds for each G such that the difference mapping $G(\cdot) - \frac{\partial F}{\partial x}(\sigma_0, x_0)$ is Lipschitz-continuous on $B_{\bar{\varepsilon}}(\bar{\xi})$ with modulus $l \in (0, \bar{l})$.

It can be easily seen that there exists $\tilde{\delta}_2 \in (0, \delta/4]$ possessing the following property: if $\eta \in \operatorname{cone} B_{\tilde{\delta}_2}(\bar{\eta}) \setminus \{0\}$ then $\|\|\bar{\eta}\|\eta/\|\eta\| - \bar{\eta}\| \le \delta/4$. Put

$$\gamma = \begin{cases} \|\bar{\eta}\| & \text{if } \bar{\eta} \ne 0, \\ \frac{\delta}{4} & \text{if } \bar{\eta} = 0. \end{cases} \tag{2.16}$$

Set $\delta_1 = \min\{\delta/16, \gamma/4\}$, $\delta_2 = \|\bar{y}\|\tilde{\delta}_2/\|\bar{\eta}\|$ if $\bar{\eta} \ne 0$ (so that $\operatorname{cone} B_{\delta_2}(\bar{y}) = \operatorname{cone} B_{\tilde{\delta}_2}(\bar{\eta})$; if $\bar{\eta} = 0$, $\delta_2 > 0$ can be taken arbitrarily). Fix $(\sigma, x) \in \Sigma \times X$ satisfying

$$F(\sigma, x) \in Q \cap B_{\delta_1}(F(\sigma_0, x_0)) - \operatorname{cone} B_{\delta_2}(\bar{y}) \tag{2.17}$$

and such that $F(\sigma, x) \notin Q$ (otherwise estimate (1.3) holds trivially). Then there exists $q = q(\sigma, x) \in Q \cap B_{\delta_1}(F(\sigma_0, x_0))$ such that

$$-(F(\sigma, x) - q) \in \operatorname{cone} B_{\tilde{\delta}_2}(\bar{\eta}),$$

and hence, according to (2.16) and to the choice of $\tilde{\delta}_2$, it holds that

$$\left\| \frac{\gamma}{\|F(\sigma, x) - q\|}(F(\sigma, x) - q) + \bar{\eta} \right\| \leq \frac{\delta}{4} \tag{2.18}$$

(note that $\|F(\sigma, x) - q\|$ cannot be equal to 0 since $F(\sigma, x) \notin Q$).

Set

$$t = t(\sigma, x, q) = \min\left\{ \frac{16 \operatorname{dist}(F(\sigma, x), Q)}{\delta}, \frac{\|F(\sigma, x) - q\|}{\gamma} \right\}. \tag{2.19}$$

Note that $t > 0$ but t tends to 0 as (σ, x) tends to (σ_0, x_0). Define the mapping $G : X \to Y$,

$$G(\xi) = G(\sigma, x; \xi) = \frac{1}{t}(F(\sigma, x + t\xi) - F(\sigma, x)), \tag{2.20}$$

and the difference mapping $\Phi : X \to Y$,

$$\begin{aligned} \Phi(\xi) &= \Phi(\sigma, x; \xi) \\ &= G(\xi) - \frac{\partial F}{\partial x}(\sigma_0, x_0)\xi \\ &= \frac{1}{t}\left(F(\sigma, x + t\xi) - F(\sigma, x) - \frac{\partial F}{\partial x}(\sigma_0, x_0)t\xi \right). \end{aligned} \tag{2.21}$$

By the mean value theorem we obtain that for (σ, x) close enough to (σ_0, x_0), and for each $\xi^1, \xi^2 \in X$,

$$\|\Phi(\xi^1) - \Phi(\xi^2)\| \leq \sup_{\theta \in [0, 1]} \left\| \frac{\partial F}{\partial x}(\sigma, x + t(\theta\xi^1 + (1 - \theta)\xi^2)) - \frac{\partial F}{\partial x}(\sigma_0, x_0) \right\| \|\xi^1 - \xi^2\|,$$

and hence, there exist a neighborhood U of σ_0 and $\varepsilon > 0$ such that Φ is Lipschitz-continuous on $B_{\bar{\varepsilon}}(\bar{\xi})$ with modulus $l \in (0, \bar{l})$ provided $(\sigma, x) \in U \times B_\varepsilon(x_0)$. Throughout the rest of the proof we suppose that the latter inclusion holds. Then by choosing another ("smaller") U and by reducing $\varepsilon > 0$ (if necessary), we obtain

$$\|\Phi(\bar{\xi})\| \leq \sup_{\theta \in [0, 1]} \left\| \frac{\partial F}{\partial x}(\sigma, x + t\theta\bar{\xi}) - \frac{\partial F}{\partial x}(\sigma_0, x_0) \right\| \|\bar{\xi}\| \leq \frac{\delta}{2}. \tag{2.22}$$

Set

$$\theta = \theta(\sigma, x, q) = \frac{2\|F(\sigma, x) - q\|}{\gamma}, \tag{2.23}$$

$$\tilde{y} = \tilde{y}(\sigma, x, q) = \theta F(\sigma_0, x_0) + (1 - \theta)q. \tag{2.24}$$

Note that, by the definition of δ_1, $\theta \in (0, 1]$ provided U and $\varepsilon > 0$ are chosen appropriately. Choose an element $p = p(\sigma, x) \in Q$ such that

$$\|F(\sigma, x) - p\| \leq 2 \operatorname{dist}(F(\sigma, x), Q), \tag{2.25}$$

and set

$$\tau = \tau(\sigma, x, q) = \frac{\gamma t}{\|F(\sigma, x) - q\|}, \tag{2.26}$$

$$y = y(\sigma, x, p, q) = -\frac{1}{t}(\tau(F(\sigma, x) - \tilde{y}) + (1 - \tau)(F(\sigma, x) - p)). \tag{2.27}$$

Note that $\tau \in (0, 1]$, and moreover, $\tau = 1$ provided

$$\|F(\sigma, x) - q\|/\gamma \le 16\,\mathrm{dist}(F(\sigma, x), Q)/\delta,$$

that is, when $t = \|F(\sigma, x) - q\|/\gamma$ (see (2.19)). Taking this into account, by (2.18), (2.19), (2.23)–(2.27), and the definition of δ_1, we derive that

$$\begin{aligned}
\|y - \bar{\eta}\| &= \left\| \frac{\tau}{t}(F(\sigma, x) - \tilde{y}) + \frac{1 - \tau}{t}(F(\sigma, x) - p) + \bar{\eta} \right\| \\
&\le \left\| \frac{\gamma}{\|F(\sigma, x) - q\|}(F(\sigma, x) - \tilde{y}) + \bar{\eta} \right\| + (1 - \tau)\frac{\|F(\sigma, x) - p\|}{t} \\
&\le \left\| \frac{\gamma}{\|F(\sigma, x) - q\|}(F(\sigma, x) - q) + \bar{\eta} \right\| \\
&\quad + \theta\frac{\gamma\|q - F(\sigma_0, x_0)\|}{\|F(\sigma, x) - q\|} + \frac{2\delta\,\mathrm{dist}(F(\sigma, x), Q)}{16\,\mathrm{dist}(F(\sigma, x), Q)} \\
&\le \frac{\delta}{4} + \frac{\delta}{8} + \frac{\delta}{8} \\
&= \frac{\delta}{2}.
\end{aligned}$$

Thus, by the second equality in (2.21), and by (2.22), it holds that

$$\begin{aligned}
\left\| y - \bar{\eta} + \frac{\partial F}{\partial x}(\sigma_0, x_0)\bar{\xi} - G(\bar{\xi}) \right\| &\le \|y - \bar{\eta}\| + \|\Phi(\bar{\xi})\| \\
&\le \frac{\delta}{2} + \frac{\delta}{2} \\
&\le \delta.
\end{aligned} \tag{2.28}$$

Hence, the estimate (2.15) must be valid for y defined in (2.27) and for G defined in (2.20) provided U and $\varepsilon > 0$ are chosen appropriately. This means that there exist $\xi = \xi(\sigma, x, p, q) \in X$ and $\eta = \eta(\sigma, x, p, q) \in Q$ such that

$$G(\xi) = y + \eta - F(\sigma_0, x_0) \tag{2.29}$$

and

$$\begin{aligned}
\|\xi\| &\le \|\bar{\xi}\| + \|\xi - \bar{\xi}\| \\
&\le \|\bar{\xi}\| + \bar{c}\,\mathrm{dist}(G(\bar{\xi}) - y, Q - F(\sigma_0, x_0)) \\
&\le \|\bar{\xi}\| + \bar{c}\|G(\bar{\xi}) - y\| \\
&\le \|\bar{\xi}\| + \bar{c}\left(\left\| \frac{\partial F}{\partial x}(\sigma_0, x_0)\bar{\xi} - \bar{\eta} \right\| + \delta \right),
\end{aligned} \tag{2.30}$$

where (2.28) and the inclusion $0 \in Q - F(\sigma_0, x_0)$ were taken into account. Note that the right-hand side of the last relation is a constant independent of σ, x, p, and q.

Employing (2.21), (2.24), (2.27), and (2.29), we have

$$\begin{aligned} F(\sigma, x + t\xi) &= t\Phi(\xi) + F(\sigma, x) + t\frac{\partial F}{\partial x}(\sigma_0, x_0)\xi \\ &= t\Phi(\xi) + t\frac{\partial F}{\partial x}(\sigma_0, x_0)\xi \\ &\quad + \tau(F(\sigma, x) - \tilde{y}) + \tau\tilde{y} + (1-\tau)(F(\sigma, x) - p) + (1-\tau)p \\ &= tG(\xi) - ty + \tau\tilde{y} + (1-\tau)p \\ &= t(\eta - F(\sigma_0, x_0)) + \tau\tilde{y} + (1-\tau)p \\ &= t\eta - tF(\sigma_0, x_0) + \tau\theta F(\sigma_0, x_0) + \tau(1-\theta)q + (1-\tau)p \\ &= t\eta + (\tau\theta - t)F(\sigma_0, x_0) + \tau(1-\theta)q + (1-\tau)p, \end{aligned}$$

where the right-hand side is a convex combination of η, $F(\sigma_0, x_0)$, p, and q provided U and $\varepsilon > 0$ are chosen appropriately. However, all the elements η, $F(\sigma_0, x_0)$, p, and q belong to the convex set Q. Hence,

$$F(\sigma, x + t\xi) \in Q,$$

and moreover, by (2.19) and (2.30),

$$t\|\xi\| \le c\,\mathrm{dist}(F(\sigma, x), Q),$$

where $c = 16(\|\bar{\xi}\| + \bar{c}(\|\frac{\partial F}{\partial x}(\sigma_0, x_0)\bar{\xi} - \bar{\eta}\| + \delta))/\delta$.

We thus proved that (1.3) holds for all $(\sigma, x) \in U \times B_\varepsilon(x_0)$ satisfying (2.17). In order to complete the proof it suffices to refer the reader to Proposition 2.2. □

3. Directional metric regularity. Let (X, ρ) be a complete metric space and Y be a normed linear space. As will be explained below, the following definition is motivated by Theorem 2.3.

DEFINITION 3.1. *The multifunction* $\Psi\colon X \to 2^Y$ *is* metrically regular at a point $(x_0, y_0) \in \mathrm{graph}\,\Psi$ in a direction $\bar{y} \in Y$, *at a rate* $c > 0$, *if there exist* $\varepsilon > 0$ *and* $\delta > 0$ *such that the estimate*

$$\mathrm{dist}(x, \Psi^{-1}(y)) \le c\,\mathrm{dist}(y, \Psi(x)) \tag{3.1}$$

holds for all $(x, y) \in B_\varepsilon(x_0) \times B_\varepsilon(y_0)$ *satisfying the inclusion*

$$y \in \Psi(x) + \mathrm{cone}\,B_\delta(\bar{y}). \tag{3.2}$$

Evidently, if $\bar{y} = 0$, then directional metric regularity turns into the usual metric regularity. Moreover, if the latter holds, directional metric regularity holds in any direction $\bar{y} \in Y$, including $\bar{y} = 0$. At the same time, directional metric regularity can hold when the usual metric regularity is violated; see Example 3.1 below.

Recall that the multifunction $\Psi : X \to 2^Y$ is said to be *lower* (or *inner*) *semicontinuous at a point* $(x_0, y_0) \in \mathrm{graph}\,\Psi$ if for any sequence $\{x^k\} \subset X$ convergent to x_0 there exists a sequence $\{y^k\} \subset Y$ convergent to y_0 such that $y^k \in \Psi(x^k)$ for all k (see, e.g., [16, Definition 1.63]).

The next theorem follows the pattern of [3, Theorem 2.84], [16, Theorem 4.25]; it says that the property of directional metric regularity of Ψ in a given direction $\bar{y}$ is

stable subject to small Lipschitzian single-valued perturbations of Ψ. For the usual notion of metric regularity, this property was studied, e.g., in [9, 7, 8]. Yet another reference to be mentioned in relation with this property is [6], where the importance of stability of regularity properties with respect to perturbations is already completely clear.

THEOREM 3.2. *Let $(x_0, y_0) \in \operatorname{graph}\Psi$. Assume that the multifunction Ψ is closed, lower semicontinuous at (x_0, y_0), and metrically regular at (x_0, y_0) in a direction $\bar{y} \in Y$, at a rate $c > 0$. Let $\varepsilon > 0$ and $\delta > 0$ be chosen according to Definition* 3.1*, and set*

$$\alpha = \begin{cases} \frac{\delta}{\|\bar{y}\|} & \text{if } \|\bar{y}\| \geq \delta, \\ +\infty & \text{if } \|\bar{y}\| < \delta, \end{cases} \qquad \beta(\alpha) = \begin{cases} \frac{2}{1+\alpha} & \text{if } \alpha < +\infty, \\ 0 & \text{if } \alpha = +\infty. \end{cases} \tag{3.3}$$

Then for any mapping $\Phi : X \to Y$ which is Lipschitz-continuous on $B_\varepsilon(x_0)$ with modulus $l > 0$ such that

$$c\,l < \min\{1,\, \alpha/5\}, \tag{3.4}$$

the multifunction $\Psi + \Phi$ is metrically regular at $(x_0, y_0 + \Phi(x_0))$ in the direction $\bar{y}$, at a rate $\tilde{c} = c(1 - c\,l)^{-1}(1 + \beta(\alpha))$.

In the case of usual metric regularity (i.e., when $\bar{y} = 0$), Theorem 3.2 directly reduces to [16, Theorem 4.25] but with an extraneous assumption of lower semicontinuity. It is possible that this assumption can actually be removed in Theorem 3.2, though the authors did not manage to avoid it.

Remark 3.1. As can be seen from the proof below, the assertion of Theorem 3.2 can be replaced by a somewhat stronger one: Under the assumptions of this theorem, for each $l > 0$ satisfying (3.4) there exist $\tilde{\varepsilon} > 0$ and $\tilde{\delta} > 0$ such that the estimate

$$\operatorname{dist}(x,\, (\Psi + \Phi)^{-1}(y)) \leq \tilde{c}\operatorname{dist}(y,\, \Psi(x) + \Phi(x)) \tag{3.5}$$

holds for any mapping $\Phi : X \to Y$ which is Lipschitz-continuous on $B_\varepsilon(x_0)$ with modulus l, and for all $(x, y) \in B_{\tilde{\varepsilon}}(x_0) \times B_{\tilde{\varepsilon}}(y_0 + \Phi(x_0))$ satisfying the inclusion

$$y \in \Psi(x) + \Phi(x) + \operatorname{cone} B_{\tilde{\delta}}(\bar{y}). \tag{3.6}$$

That is, $\tilde{\varepsilon}$ and $\tilde{\delta}$ do not depend on a specific Φ but only on ε, δ, c, $\|\bar{y}\|$, and l.

Proof. Let $\varepsilon > 0$ and $\delta > 0$ be chosen according to Definition 3.1. Fix arbitrary $\tilde{\varepsilon} \in (0,\, \varepsilon]$, $\tilde{\delta} \in (0,\, \delta)$, and $\hat{\varepsilon} > 0$ satisfying the following set of conditions:

$$\tilde{\varepsilon} + \frac{\gamma(\hat{\varepsilon})}{l}(1 - \gamma(\hat{\varepsilon}))^{-1}(1 + \beta(\alpha))(\tilde{\varepsilon} + l\tilde{\varepsilon} + \omega(\tilde{\varepsilon})) < \varepsilon, \tag{3.7}$$

$$\tilde{\varepsilon} + l\tilde{\varepsilon} + \left(\gamma(\hat{\varepsilon})(1 - \gamma(\hat{\varepsilon}))^{-1}(1 + \beta(\alpha)) + \frac{\beta(\alpha)}{2}\right)(\tilde{\varepsilon} + l\tilde{\varepsilon} + \omega(\tilde{\varepsilon})) < \varepsilon, \tag{3.8}$$

$$\gamma(\hat{\varepsilon}) < \min\left\{1,\, \frac{\alpha(\delta - \tilde{\delta})}{5\delta}\right\}, \tag{3.9}$$

where $\omega(\tilde{\varepsilon}) = \sup_{x \in B_{\tilde{\varepsilon}}(x_0)} \operatorname{dist}(y_0,\, \Psi(x))$, $\gamma(\hat{\varepsilon}) = c\,l(1 + \hat{\varepsilon})$ (note that $\omega(\tilde{\varepsilon}) \to 0$ as $\tilde{\varepsilon} \to 0$ because of the lower semicontinuity Ψ at (x_0, y_0), and recall (3.4)).

Let $(x, y) \in B_{\tilde{\varepsilon}}(x_0) \times B_{\tilde{\varepsilon}}(y_0 + \Phi(x_0))$ satisfying (3.6) be fixed. In order to prove estimate (3.5) it suffices to establish the existence of $\chi(x) \in (\Psi + \Phi)^{-1}(y)$ such that

$$\rho(x, \chi(x)) \le c(1 - c\,l)^{-1}(1 + \beta(\alpha))\,\mathrm{dist}(y, \Psi(x) + \Phi(x)). \tag{3.10}$$

The needed point $\chi(x)$ will be defined by means of the auxiliary iterative process. For that purpose set $t = t(x, y) = \mathrm{dist}(y, \Psi(x) + \Phi(x))$ and define the sequence $\{\tau_k\} \subset \mathbf{R}_+$ by setting

$$\tau_1 = \begin{cases} \frac{\delta - \tilde{\delta}}{(\|\bar{y}\| + \delta)\delta} t & \text{if } \|\bar{y}\| \ge \delta, \\ 0 & \text{if } \|\bar{y}\| < \delta, \end{cases} \qquad \tau_{k+1} = \frac{2}{5}\tau_k,\ k = 1, 2, \ldots .$$

According to (3.3),

$$\tau_1 \|\bar{y}\| \le \frac{\beta(\alpha)}{2} t. \tag{3.11}$$

Note that by the definition of $\omega(\tilde{\varepsilon})$

$$\begin{aligned} t &= \mathrm{dist}(y - \Phi(x_0) + \Phi(x_0) - \Phi(x), \Psi(x)) \\ &\le \|y - y_0 - \Phi(x_0)\| + \|\Phi(x) - \Phi(x_0)\| + \mathrm{dist}(y_0, \Psi(x)) \\ &\le \tilde{\varepsilon} + l\tilde{\varepsilon} + \omega(\tilde{\varepsilon}). \end{aligned} \tag{3.12}$$

We shall construct a sequence $\{x^k\} \subset X$ such that $x^1 = x$ and for all $k = 1, 2, \ldots$

$$y - \Phi(x^k) - \tau_k \bar{y} \in \Psi(x^{k+1}), \tag{3.13}$$

$$\rho(x^{k+2}, x^{k+1}) \le \gamma(\hat{\varepsilon})\rho(x^{k+1}, x^k) + \frac{\gamma(\hat{\varepsilon})}{l}(\tau_k - \tau_{k+1})\|\bar{y}\|, \tag{3.14}$$

$$\rho(x^{k+1}, x_0) \le \tilde{\varepsilon} + \frac{\gamma(\hat{\varepsilon})}{l}(1 - \gamma(\hat{\varepsilon}))^{-1}(1 + \beta(\alpha))(\tilde{\varepsilon} + l\tilde{\varepsilon} + \omega(\tilde{\varepsilon})), \tag{3.15}$$

and if $\|\bar{y}\| \ge \delta$ then

$$\|\Phi(x^{k+1}) - \Phi(x^k)\| \le \delta(\tau_k - \tau_{k+1}). \tag{3.16}$$

By (3.6) we obtain the existence of $\theta \ge 0$ and $\eta \in Y$ such that $\|\eta\| \le \tilde{\delta}$ and

$$y - \Phi(x^1) - \theta(\bar{y} + \eta) \in \Psi(x^1). \tag{3.17}$$

Note that if $\theta = 0$ then $x = x^1 \in (\Psi + \Phi)^{-1}(y)$, and we are done. Thus, let $\theta > 0$.

Set

$$y^1 = y - \Phi(x^1) - \tau_1\bar{y}, \quad \eta_0 = \theta(\bar{y} + \eta), \quad \eta^1 = \eta_0 - \tau_1\bar{y}.$$

By (3.8), (3.11), and (3.12) we then derive

$$\begin{aligned} \|y^1 - y_0\| &= \|y^1 + \Phi(x_0) - y_0 - \Phi(x_0)\| \\ &\le \|y - y_0 - \Phi(x_0)\| + \|\Phi(x_0) - \Phi(x^1)\| + \tau_1\|\bar{y}\| \\ &\le \tilde{\varepsilon} + l\tilde{\varepsilon} + \frac{\beta(\alpha)}{2}(\tilde{\varepsilon} + l\tilde{\varepsilon} + \omega(\tilde{\varepsilon})) \\ &< \varepsilon. \end{aligned} \tag{3.18}$$

Furthermore,

$$\begin{aligned}\|\eta_0\| &\le \theta(\|\bar{y}\| + \tilde{\delta}) \\ &< \theta(\|\bar{y}\| + \delta),\end{aligned}$$

and hence,

$$\begin{aligned}\theta &> \frac{\|\eta_0\|}{\|\bar{y}\| + \delta} \\ &\ge \frac{t}{\|\bar{y}\| + \delta} \\ &\ge \tau_1,\end{aligned} \tag{3.19}$$

where it was taken into account that, by (3.17), $y - \eta_0 \in \Psi(x^1) + \Phi(x^1)$, and hence, by the definition of t, it holds that $t \le \|\eta_0\|$. From (3.19) (including the intermediate inequalities) and the definition of τ_1 it follows that

$$\begin{aligned}\frac{\theta\|\eta\|}{\theta - \tau_1} &\le \frac{\theta\tilde{\delta}}{\theta - \frac{\delta - \tilde{\delta}}{(\|\bar{y}\| + \delta)\delta} t} \\ &\le \frac{\theta\tilde{\delta}}{\theta - \theta\frac{\delta - \tilde{\delta}}{\delta}} \\ &= \delta,\end{aligned}$$

and hence, by (3.17),

$$\begin{aligned}\eta^1 &= (\theta - \tau_1)\bar{y} + \theta\eta \\ &= (\theta - \tau_1)\left(\bar{y} + \frac{\theta\eta}{\theta - \tau_1}\right) \\ &\in \operatorname{cone} B_\delta(\bar{y}).\end{aligned} \tag{3.20}$$

Taking into account the equality $y^1 = y - \Phi(x^1) - \theta(\bar{y} + \eta) + \eta^1$, we conclude by (3.17) and (3.20) that

$$y^1 \in \Psi(x^1) + \operatorname{cone} B_\delta(\bar{y});$$

that is, (3.2) holds with (x, y) replaced by $(x^1, y^1) \in B_\varepsilon(x_0) \times B_\varepsilon(y_0)$ (see (3.18)). Thus, by metric regularity of Ψ at a point (x_0, y_0) in a direction $\bar{y}$, there exists $x^2 \in X$ such that

$$y - \Phi(x^1) - \tau_1\bar{y} \in \Psi(x^2), \tag{3.21}$$

$$\begin{aligned}\rho(x^2, x^1) &\le c(1 + \hat{\varepsilon}) \operatorname{dist}(y - \Phi(x^1) - \tau_1\bar{y}, \Psi(x^1)) \\ &\le c(1 + \hat{\varepsilon})(t + \tau_1\|\bar{y}\|) \\ &\le \frac{\gamma(\hat{\varepsilon})}{l}\left(1 + \frac{\beta(\alpha)}{2}\right) t,\end{aligned} \tag{3.22}$$

where the definition of t and (3.11) were taken into account. In particular, (3.13) holds for $k = 1$.

Employing (3.7), (3.12), and (3.22), we derive

$$\begin{aligned}\rho(x^2, x_0) &\le \rho(x^1, x_0) + \rho(x^2, x^1)\\ &\le \tilde{\varepsilon} + \frac{\gamma(\hat{\varepsilon})}{l}\left(1 + \frac{\beta(\alpha)}{2}\right) t\\ &\le \tilde{\varepsilon} + \frac{\gamma(\hat{\varepsilon})}{l}\left(1 + \frac{\beta(\alpha)}{2}\right)(\tilde{\varepsilon} + l\tilde{\varepsilon} + \omega(\tilde{\varepsilon}))\\ &< \varepsilon,\end{aligned} \tag{3.23}$$

and, in particular, (3.15) holds for $k = 1$.

Furthermore, if $\|\bar{y}\| \ge \delta$, then by (3.3), (3.9), (3.22), and (3.23), we derive

$$\begin{aligned}\|\Phi(x^2) - \Phi(x^1)\| &\le \gamma(\hat{\varepsilon})\left(1 + \frac{\beta(\alpha)}{2}\right) t\\ &< \frac{\alpha(\delta - \tilde{\delta})}{5\delta}\left(1 + \frac{1}{1+\alpha}\right) t\\ &= \frac{\alpha(\delta - \tilde{\delta})(2+\alpha)}{5(1+\alpha)\delta} t\\ &\le \frac{3\alpha(\delta - \tilde{\delta})}{5(1+\alpha)\delta} t\\ &= \delta(\tau_1 - \tau_2);\end{aligned} \tag{3.24}$$

that is, (3.16) holds for $k = 1$.

Set

$$q^2 = y - \Phi(x^1) - \tau_1 \bar{y},$$

$$y^2 = y - \Phi(x^2) - \tau_2 \bar{y}, \quad \eta^2 = (\tau_1 - \tau_2)\bar{y} + \Phi(x^1) - \Phi(x^2).$$

Note that $q^2 \in \Psi(x^2)$ by (3.21), and by (3.24) we conclude that $\eta^2 \in \operatorname{cone} B_\delta(\bar{y})$ ((3.24) holds only if $\|\bar{y}\| \ge \delta$, but otherwise, $\operatorname{cone} B_\delta(\bar{y}) = Y$).

By (3.8), (3.11), and (3.12), and by (3.15) (for $k = 1$), it follows that

$$\begin{aligned}\|y^2 - y_0\| &= \|y - y_0 - \Phi(x_0)\| + \|\Phi(x^2) - \Phi(x_0)\| + \tau_2\|\bar{y}\|\\ &\le \tilde{\varepsilon} + l\tilde{\varepsilon} + \gamma(\hat{\varepsilon})(1 - \gamma(\hat{\varepsilon}))^{-1}(1 + \beta(\alpha))(\tilde{\varepsilon} + l\tilde{\varepsilon} + \omega(\tilde{\varepsilon}))\\ &\quad + \frac{\beta(\alpha)}{2}(\tilde{\varepsilon} + l\tilde{\varepsilon} + \omega(\tilde{\varepsilon}))\\ &< \varepsilon.\end{aligned} \tag{3.25}$$

The inclusions $q^2 \in \Psi(x^2)$ and $\eta^2 \in \operatorname{cone} B_\delta(\bar{y})$ imply that

$$\begin{aligned}y^2 &= y - \Phi(x^1) - \tau_1\bar{y} + \Phi(x^1) + \tau_1\bar{y} - \Phi(x^2) - \tau_2\bar{y}\\ &= q^2 + \eta^2\\ &\in \Psi(x^2) + \operatorname{cone} B_\delta(\bar{y});\end{aligned}$$

that is, (3.2) holds with (x, y) replaced by $(x^2, y^2) \in B_\varepsilon(x_0) \times B_\varepsilon(y_0)$ (see (3.23), (3.25)). Thus, by metric regularity of Ψ at a point (x_0, y_0) in a direction $\bar{y}$, there exists $x^3 \in X$ such that

$$y - \Phi(x^2) - \tau_2\bar{y} \in \Psi(x^3), \tag{3.26}$$

$$
\begin{aligned}
\rho(x^3,\,x^2) &\le c(1+\hat\varepsilon)\,\mathrm{dist}(y-\Phi(x^2)-\tau_2\bar y,\,\Psi(x^2))\\
&\le c(1+\hat\varepsilon)\|y-\Phi(x^2)-\tau_2\bar y-q^2\|\\
&\le c(1+\hat\varepsilon)(\|\Phi(x^2)-\Phi(x^1)\|+(\tau_1-\tau_2)\|\bar y\|)\\
&\le \gamma(\hat\varepsilon)\rho(x^2,\,x^1)+\frac{\gamma(\hat\varepsilon)}{l}(\tau_1-\tau_2)\|\bar y\|,
\end{aligned}
\tag{3.27}
$$

where the definition of q^2 was taken into account. In particular, (3.13) holds for $k=2$, and (3.14) holds for $k=1$.

Employing (3.11), (3.22), and (3.27), we derive

$$
\begin{aligned}
\rho(x^3,\,x^1) &\le \rho(x^2,\,x^1)+\rho(x^3,\,x^2)\\
&\le (1+\gamma(\hat\varepsilon))\rho(x^2,\,x^1)+\frac{\gamma(\hat\varepsilon)}{l}(\tau_1-\tau_2)\|\bar y\|\\
&\le (1+\gamma(\hat\varepsilon))\left(\rho(x^2,\,x^1)+\frac{\gamma(\hat\varepsilon)}{l}\tau_1\|\bar y\|\right)\\
&\le \frac{\gamma(\hat\varepsilon)}{l}(1+\gamma(\hat\varepsilon))\left(\left(1+\frac{\beta(\alpha)}{2}\right)t+\tau_1\|\bar y\|\right)\\
&\le \frac{\gamma(\hat\varepsilon)}{l}(1+\gamma(\hat\varepsilon))(1+\beta(\alpha))t,
\end{aligned}
$$

and thus, by (3.7) and (3.12),

$$
\begin{aligned}
\rho(x^3,\,x_0) &\le \rho(x^1,\,x_0)+\rho(x^3,\,x^1)\\
&\le \tilde\varepsilon+\frac{\gamma(\hat\varepsilon)}{l}(1+\gamma(\hat\varepsilon))(1+\beta(\alpha))t\\
&\le \tilde\varepsilon+\frac{\gamma(\hat\varepsilon)}{l}(1-\gamma(\hat\varepsilon))^{-1}(1+\beta(\alpha))(\tilde\varepsilon+l\tilde\varepsilon+\omega(\tilde\varepsilon))\\
&< \varepsilon,
\end{aligned}
\tag{3.28}
$$

where the evident inequality $1+\gamma(\hat\varepsilon)<(1-\gamma(\hat\varepsilon))^{-1}$ was also employed. In particular, (3.15) holds for $k=2$.

Furthermore, if $\|\bar y\|\ge\delta$, then by (3.3), (3.9), (3.24), and (3.28), and by the intermediate inequalities in (3.27), we derive

$$
\begin{aligned}
\|\Phi(x^3)-\Phi(x^2)\| &\le l\rho(x^3,\,x^2)\\
&\le \gamma(\hat\varepsilon)(\|\Phi(x^2)-\Phi(x^1)\|+(\tau_2-\tau_1)\|\bar y\|)\\
&\le \gamma(\hat\varepsilon)(\delta+\|\bar y\|)(\tau_1-\tau_2)\\
&< \frac{\alpha}{5}(\delta+\|\bar y\|)\frac{5}{2}\left(1-\frac{2}{5}\right)\tau_2\\
&= \frac{\delta}{2\|\bar y\|}(\delta+\|\bar y\|)(\tau_2-\tau_3)\\
&\le \delta(\tau_2-\tau_3);
\end{aligned}
$$

that is, (3.16) holds for $k=2$.

Suppose now that for some $s\ge 3$ we have already constructed points $x^k\in X$, $k=1,\dots,s$, such that (3.13), (3.15), and (3.16), if $\|\bar y\|\ge\delta$, hold for each $k=1,\dots,s-1$, and (3.14) holds for each $k=1,\dots,s-2$. Set

$$
q^s=y-\Phi(x^{s-1})-\tau_{s-1}\bar y,
$$

$$y^s = y - \Phi(x^s) - \tau_s\bar{y}, \quad \eta^s = (\tau_{s-1} - \tau_s)\bar{y} + \Phi(x^{s-1}) - \Phi(x^s).$$

Note that $q^s \in \Phi(x^s)$ by (3.13) (with $k = s-1$), and by (3.16) (with $k = s-1$) we conclude that $\eta^s \in \operatorname{cone} B_\delta(\bar{y})$ ((3.16) holds only if $\|\bar{y}\| \geq \delta$, but otherwise, $\operatorname{cone} B_\delta(\bar{y}) = Y$).

By (3.8), (3.11), and (3.12), and by (3.15) (for $k = s-1$), it follows that

$$\begin{aligned}(3.29)\qquad \|y^s - y_0\| &= \|y - y_0 - \Phi(x_0)\| + \|\Phi(x^s) - \Phi(x_0)\| + \tau_s\|\bar{y}\| \\ &\leq \tilde{\varepsilon} + l\tilde{\varepsilon} + \gamma(\hat{\varepsilon})(1-\gamma(\hat{\varepsilon}))^{-1}(1+\beta(\alpha))(\tilde{\varepsilon} + l\tilde{\varepsilon} + \omega(\tilde{\varepsilon})) \\ &\quad + \frac{\beta(\alpha)}{2}(\tilde{\varepsilon} + l\tilde{\varepsilon} + \omega(\tilde{\varepsilon})) \\ &< \varepsilon.\end{aligned}$$

The inclusions $q^s \in \Phi(x^s)$ and $\eta^s \in \operatorname{cone} B_\delta(\bar{y})$ imply that

$$\begin{aligned} y^s &= y - \Phi(x^{s-1}) - \tau_{s-1}\bar{y} + \Phi(x^{s-1}) + \tau_{s-1}\bar{y} - \Phi(x^s) - \tau_s\bar{y} \\ &= q^s + \eta^s \\ &\in \Psi(x^s) + \operatorname{cone} B_\delta(\bar{y}); \end{aligned}$$

that is, (3.2) holds with (x, y) replaced by $(x^s, y^s) \in B_\varepsilon(x_0) \times B_\varepsilon(y_0)$ (see (3.8), (3.15), (3.29)). Thus, by metric regularity of Ψ at a point (x_0, y_0) in a direction $\bar{y}$, there exists $x^{s+1} \in X$ such that

$$y - \Phi(x^s) - \tau_s\bar{y} \in \Psi(x^{s+1}),$$

$$\begin{aligned}(3.30)\qquad \rho(x^{s+1}, x^s) &\leq c(1+\hat{\varepsilon})\operatorname{dist}(y - \Phi(x^s) - \tau_s\bar{y}, \Psi(x^s)) \\ &\leq c(1+\hat{\varepsilon})\|y - \Phi(x^s) - \tau_s\bar{y} - q^s\| \\ &\leq c(1+\hat{\varepsilon})(\|\Phi(x^s) - \Phi(x^{s-1})\| + (\tau_{s-1} - \tau_s)\|\bar{y}\|) \\ &\leq \gamma(\hat{\varepsilon})\rho(x^s, x^{s-1}) + \frac{\gamma(\hat{\varepsilon})}{l}(\tau_{s-1} - \tau_s)\|\bar{y}\|,\end{aligned}$$

where the definition of q^s was taken into account. In particular, (3.13) holds for $k = s$, and (3.14) holds for $k = s-1$.

Employing (3.14) we derive that for each $k = 1, \ldots, s-1$

$$\begin{aligned}\sum_{i=1}^{k}\rho(x^{i+2}, x^{i+1}) &\leq \sum_{i=1}^{k}\left(\gamma(\hat{\varepsilon})\rho(x^{i+1}, x^i) + \frac{\gamma(\hat{\varepsilon})}{l}(\tau_i - \tau_{i+1})\|\bar{y}\|\right) \\ &\leq \gamma(\hat{\varepsilon})\sum_{i=1}^{k}\rho(x^{i+1}, x^i) + \frac{\gamma(\hat{\varepsilon})}{l}\tau_1\|\bar{y}\|.\end{aligned}$$

It can be easily seen by induction that the latter property implies the estimate

$$(3.31)\qquad \sum_{k=1}^{s-1}\rho(x^{k+2}, x^{k+1}) \leq \gamma(\hat{\varepsilon})(1-\gamma(\hat{\varepsilon}))^{-1}(\rho(x^2, x^1) + l^{-1}\tau_1\|\bar{y}\|).$$

Hence, by (3.11), (3.22),

$$\begin{aligned}(3.32)\ \rho(x^{s+1}, x^1) &\leq \rho(x^2, x^1) + \sum_{i=1}^{s-1}\rho(x^{i+2}, x^{i+1}) \\ &\leq \left(1 + \gamma(\hat{\varepsilon})(1-\gamma(\hat{\varepsilon}))^{-1}\right)\rho(x^2, x^1) + \frac{\gamma(\hat{\varepsilon})}{l}(1-\gamma(\hat{\varepsilon}))^{-1}\tau_1\|\bar{y}\| \\ &\leq \frac{\gamma(\hat{\varepsilon})}{l}(1-\gamma(\hat{\varepsilon}))^{-1}(1+\beta(\alpha))t,\end{aligned}$$

and thus, by (3.7) and (3.12),

$$\begin{aligned}\rho(x^{s+1}, x_0) &\le \rho(x^1, x_0) + \rho(x^{s+1}, x^1)\\ &\le \tilde{\varepsilon} + \frac{\gamma(\hat{\varepsilon})}{l}(1-\gamma(\hat{\varepsilon}))^{-1}(1+\beta(\alpha))t\\ &\le \tilde{\varepsilon} + \frac{\gamma(\hat{\varepsilon})}{l}(1-\gamma(\hat{\varepsilon}))^{-1}(1+\beta(\alpha))(\tilde{\varepsilon} + l\tilde{\varepsilon} + \omega(\tilde{\varepsilon}))\\ &< \varepsilon.\end{aligned} \tag{3.33}$$

In particular, (3.15) holds for $k = s$.

Finally, if $\|\bar{y}\| \ge \delta$, then by (3.3), (3.9), (3.16) (with $k = s-1$), and (3.33), and by the intermediate inequalities in (3.30), we derive

$$\begin{aligned}\|\Phi(x^{s+1}) - \Phi(x^s)\| &\le l\rho(x^{s+1}, x^s)\\ &\le \gamma(\hat{\varepsilon})(\|\Phi(x^s) - \Phi(x^{s-1})\| + (\tau_{s-1} - \tau_s)\|\bar{y}\|)\\ &\le \gamma(\hat{\varepsilon})(\delta + \|\bar{y}\|)(\tau_{s-1} - \tau_s)\\ &< \frac{\alpha}{5}(\delta + \|\bar{y}\|)\frac{5}{2}\left(1 - \frac{2}{5}\right)\tau_s\\ &= \frac{\delta}{2\|\bar{y}\|}(\delta + \|\bar{y}\|)(\tau_s - \tau_{s+1})\\ &\le \delta(\tau_s - \tau_{s+1});\end{aligned}$$

that is, (3.16) holds for $k = s$.

The sequence $\{x^k\}$ with the needed properties is thus constructed. Moreover, as was shown above, (3.14) implies that (3.31) and (3.32) hold for each $s = 2, 3, \ldots$. Clearly, (3.31) implies that $\{x^k\}$ is a Cauchy sequence, and by completeness of the metric space (X, ρ), this sequence converges to some element $\chi(x) \in B_\varepsilon(x_0)$, where the last inclusion follows from (3.7) and (3.15). Since $\tau_k \to 0$ as $k \to \infty$, by passing to the limit in (3.13) we conclude that $\chi(x) \in (\Psi + \Phi)^{-1}(y)$, where closedness Ψ and continuity of Φ on $B_\varepsilon(x_0)$ were taken into account. Finally, since $\hat{\varepsilon} > 0$ can be taken arbitrarily small, (3.32) implies (3.10). □

The set $\operatorname{cone} B_\delta(\bar{y})$ in the right-hand side of (3.2) can be regarded as a *conic neighborhood* of $\bar{y}$. Note that α defined in (3.3) is invariant with respect to the choice of specific $\bar{y}$ and δ defining the same conic neighborhood, and it is natural to refer to this quantity as the *radius* of the conic neighborhood in question.

We now turn our attention to the multifunctions of the form $\Psi(x) = \Psi_F(x) = F(x) - Q$, where $F : X \to Y$ is a given mapping and $Q \subset Y$ is a given set. Note that if F is continuous at x_0 then this multifunction is automatically lower semicontinuous at (x_0, y_0) for any $y_0 \in \Psi(x_0)$. Being applied to such a multifunction, estimate (3.1) takes the form (1.4), while condition (3.2) takes the form

$$F(x) - y \in Q - \operatorname{cone} B_\delta(\bar{y}). \tag{3.34}$$

Definition 3.1 applied to $\Psi = \Psi_F$ and $y_0 = 0$ takes the following form.

DEFINITION 3.3. *The mapping $F : X \to Y$ is* metrically regular at $x_0 \in F^{-1}(Q)$ with respect to Q in a direction $\bar{y} \in Y$, *at a rate $c > 0$, if there exist $\varepsilon > 0$ and $\delta > 0$ such that the estimate* (1.4) *holds for all $(x, y) \in B_\varepsilon(x_0) \times B_\varepsilon(0)$ satisfying the inclusion* (3.34).

Throughout the rest of the paper let X and Y be Banach spaces. For a nonparametric mapping F, the directional regularity condition in a direction $\bar{y}$ takes the

form

$$0 \in \operatorname{int}(F(x_0) + \operatorname{im} F'(x_0) - \operatorname{cone}\{\bar{y}\} - Q), \tag{3.35}$$

and if Q is closed and convex then according to Theorem 2.3 (applied to $F(\sigma, x) = F(x) - \sigma$, $\sigma = y$), under the appropriate smoothness assumptions, the latter condition implies metric regularity in a direction $\bar{y}$. The converse implication can be derived from Theorem 3.2, which results in the following proposition.

PROPOSITION 3.4. *Let Q be closed and convex, and let $x_0 \in F^{-1}(Q)$. Let F be Fréchet-differentiable near x_0, and let its derivative be continuous at x_0.*

Then F is metrically regular at x_0 with respect to Q in a direction $\bar{y} \in Y$ if and only if it is regular at x_0 in this direction.

Proof. Let F be metrically regular at x_0 with respect to Q in a direction $\bar{y} \in Y$, at a rate $c > 0$. Define the mapping $\Phi : X \to Y$:

$$\Phi(x) = F(x_0) + F'(x_0)(x - x_0) - F(x).$$

Then $F + \Phi$ is a linearization of F at x_0. By the mean value theorem, for all x^1, $x^2 \in X$ close enough to x_0 we obtain

$$\begin{aligned}\|\Phi(x^1) - \Phi(x^2)\| &= \|F(x^1) - F(x^2) - F'(x_0)(x^1 - x^2)\| \\ &\le \sup_{\theta \in [0,\,1]} \|F'(\theta x^1 + (1-\theta)x^2) - F'(x_0)\| \|x^1 - x^2\|,\end{aligned}$$

and hence, Φ is Lipschitz-continuous near x_0 with modulus l, with $l > 0$ as small as needed. Applying Theorem 3.2 to $\Psi = \Psi_F$, we conclude that the linearized mapping $F + \Phi$ is metrically regular at x_0 with respect to Q in a direction $\bar{y} \in Y$, at some rate $\tilde{c} > 0$ (note that $\Psi_F + \Phi = \Psi_{F+\Phi}$). This means that there exist $\tilde{\varepsilon} > 0$ and $\tilde{\delta} > 0$ such that the estimate

$$\begin{aligned}\operatorname{dist}(x, x_0 + (F'(x_0))^{-1}(Q + y - F(x_0))) \le \tilde{c} \operatorname{dist}(&F(x_0) \\ &+ F'(x_0)(x - x_0) - y, Q)\end{aligned} \tag{3.36}$$

holds for $(x, y) \in B_{\tilde{\varepsilon}}(x_0) \times B_{\tilde{\varepsilon}}(0)$ satisfying the inclusion

$$F(x_0) + F'(x_0)(x - x_0) - y \in Q - \operatorname{cone} B_{\tilde{\delta}}(\bar{y}). \tag{3.37}$$

Take $x = x_0$, $y = -\theta\eta$, where $\eta \in B_{\tilde{\delta}}(\bar{y})$ and $\theta \ge 0$. Then (3.37) is evidently satisfied, and $y \in B_{\tilde{\varepsilon}}(0)$ for all $\theta > 0$ small enough (specifically, for all $\theta \in (0, \tilde{\varepsilon}/(\|\bar{y}\| + \tilde{\delta}))$). Hence, (3.36) holds for chosen x and y, which implies that for all $\eta \in B_{\tilde{\delta}}(\bar{y})$ and all $\theta > 0$ small enough

$$(F'(x_0))^{-1}(Q + \theta\eta - F(x_0)) \ne \emptyset,$$

and hence, there exist $\xi \in X$ and $q \in Q$ such that

$$F'(x_0)\xi = q + \theta\eta - F(x_0),$$

i.e.,

$$\theta\eta \in F(x_0) + \operatorname{im} F'(x_0) - Q.$$

It follows that

$$B_{\tilde{\delta}}(\bar{y}) \subset \operatorname{im} F'(x_0) - R_Q(F(x_0)).$$

It remains to employ Proposition 2.1 (see (2.2)). □

As mentioned above, directional metric regularity can hold when the usual metric regularity is violated. Moreover, let, e.g., $\operatorname{int} Q \neq \emptyset$ (which in particular covers the case of finitely many inequality constraints). It can be shown that in this case directional regularity condition (1.5), and hence, the directional metric regularity condition hold in any direction $\bar{y} \in -\operatorname{int} R_Q(F(\sigma_0, x_0)) \neq \emptyset$.

Example 3.1. Let $X = Y = \mathbf{R}^2$, $F(x) = (x_1, x_1^2 - x_2^2)$, $Q = \mathbf{R}^2_+$. Robinson's CQ does not hold at $x_0 = 0$, and hence, the mapping F is not metrically regular at x_0. Moreover, estimate (1.4) does not hold even on the subspaces $\{x_0\} \times Y$ and $X \times \{0\}$. Indeed, if, e.g., $y = (0, y_2)$ with $y_2 < 0$, it holds that $\operatorname{dist}(x_0, F^{-1}(Q-y)) = (-y_2)^{1/2}$, and the estimate (1.4) does not hold even for $x = x_0$. Moreover, if, e.g., $x = (0, x_2)$ with $x_2 \neq 0$, then $\operatorname{dist}(x, F^{-1}(Q)) = |x_2|/\sqrt{2}$, while $\operatorname{dist}(F(x), Q) = x_2^2$, and the estimate (1.4) does not hold even for $y = 0$.

At the same time, directional regularity condition (3.35) holds at x_0 in any direction $\bar{y} \in \mathbf{R}^2$ with $\bar{y}_2 < 0$, and hence, F is metrically regular at x_0 in each such direction.

To complete this section we note that Theorem 2.3 can actually be derived from a "uniform version" of Theorem 3.2, following the line of the argument in [3, pp. 63, 64], justifying Robinson's stability theorem.

4. Applications to sensitivity theory. Let Σ be a normed linear space. As an application of Theorem 2.3, we next show how it can be used in order to directly (that is, without employing any additional tools, with the only exception for the mean value theorem) obtain some principal lemmas playing the crucial role in sensitivity analysis under the more special directional regularity condition (1.6). We emphasize that both results presented below are known; the difference is only in the proofs. The first result is [3, Lemma 4.10].

LEMMA 4.1. *Let Q be closed and convex, and let $x_0 \in D(\sigma_0)$. Let F possess the Lipschitz-continuous derivative near (σ_0, x_0).*

If (1.6) *holds at x_0 with respect to a direction $d \in \Sigma$, then there exist $\bar{t} > 0$, $\varepsilon_1 > 0$, $\varepsilon_2 > 0$, and $a > 0$ possessing the following property: for any mappings $\rho(\cdot) : \mathbf{R}_+ \to \Sigma$ and $x(\cdot) : \mathbf{R}_+ \to X$ such that $\rho(t) = o(t)$ and the estimates*

$$\|x(t) - x_0\| \leq \varepsilon_1 t^{1/2} \tag{4.1}$$

and

$$\operatorname{dist}(F(\sigma_0 + td + \rho(t),\, x(t)),\, Q) \leq \varepsilon_2 t \tag{4.2}$$

hold for all $t \geq 0$ small enough, the estimate

(4.3)

$$\operatorname{dist}(x(t),\, D(\sigma_0 + td + \rho(t))) \leq a \left(1 + \frac{\|x(t) - x_0\|}{t}\right) \operatorname{dist}(F(\sigma_0 + td + \rho(t),\, x(t)),\, Q)$$

holds for all $t \in (0, \bar{t}]$.

Proof. As was already mentioned in section 1, (1.6) precisely coincides with (1.5) with $\bar{y} = -\frac{\partial F}{\partial \sigma}(\sigma_0, x_0)d$. For this $\bar{y}$, define $\varepsilon > 0$, $\delta > 0$, and $c > 0$ according to Theorem 2.3. Let $l > 0$ stand for the Lipschitz constant of F and $L > 0$ stand for the Lipschitz constant for the derivative of F on $B_\varepsilon(\sigma_0) \times B_\varepsilon(x_0)$ (ε can be reduced,

if necessary). For each $t > 0$ put $\sigma(t) = \sigma_0 + td + \rho(t)$. Set $\varepsilon_1 = (\delta/6L)^{1/2}$ and $\varepsilon_2 = \delta/12$, and choose $\bar{t} > 0$ such that for all $t \in (0, \bar{t}]$

$$\delta t^{1/2} \le \varepsilon, \quad \|td + \rho(t)\| \le \varepsilon, \tag{4.4}$$

$$L\left(\frac{\|td + \rho(t)\|^2}{t} + 2\varepsilon_1 \left\|d + \frac{\rho(t)}{t}\right\| t^{1/2}\right) \le \frac{\delta}{6}, \tag{4.5}$$

$$\left\|\frac{\partial F}{\partial \sigma}(\sigma_0, x_0)\right\| \frac{\|\rho(t)\|}{t} \le \frac{\delta}{6}. \tag{4.6}$$

For each $t > 0$ put

$$\tau(t) = \frac{12\,\mathrm{dist}(F(\sigma(t), x(t)), Q)}{\delta t}, \tag{4.7}$$

$$\tilde{x}(t) = \tau(t)x_0 + (1 - \tau(t))x(t), \tag{4.8}$$

$$\Phi_1(t) = F(\sigma(t), \tilde{x}(t)) - F(\sigma(t), x(t)) + \tau(t)\frac{\partial F}{\partial x}(\sigma_0, x_0)(x(t) - x_0), \tag{4.9}$$

$$\begin{aligned}\Phi_2(t) = F(\sigma(t), x(t)) - F(\sigma_0, x_0) - \frac{\partial F}{\partial \sigma}(\sigma_0, x_0)(td + \rho(t)) \\ - \frac{\partial F}{\partial x}(\sigma_0, x_0)(x(t) - x_0),\end{aligned} \tag{4.10}$$

choose an element $p(t) \in Q$ such that

$$\|F(\sigma(t), x(t)) - p(t)\| \le 2\,\mathrm{dist}(F(\sigma(t), x(t)), Q), \tag{4.11}$$

and set

$$q(t) = \tau(t)F(\sigma_0, x_0) + (1 - \tau(t))p(t). \tag{4.12}$$

Throughout the rest of the proof we assume that $F(\sigma(t), x(t)) \notin Q$ (otherwise estimate (4.3) holds trivially). Then according to (4.2), (4.7), and the definition of ε_2, it holds that

$$0 < \tau(t) = \frac{12\,\mathrm{dist}(F(\sigma(t), x(t)), Q)}{\delta t} \le \frac{12\varepsilon_2}{\delta} = 1.$$

In particular, by (4.12), $q(t) \in Q$. Furthermore, by (4.2), (4.4), and (4.8) it holds that $\sigma(t) \in B_\varepsilon(\sigma_0)$, $\tilde{x}(t) \in B_\varepsilon(x_0)$.

We next estimate $\|\Phi_1(t)\|$ and $\|\Phi_2(t)\|$ for $t \in (0, \bar{t}]$. By (4.1), (4.5), (4.8), (4.9),

the mean value theorem, and the definition of ε_1, we obtain

(4.13)
$$\begin{aligned}
\|\Phi_1(t)\| &= \Big\| F(\sigma(t),\, x(t) - \tau(t)(x(t) - x_0)) - F(\sigma(t),\, x(t)) \\
&\quad - \frac{\partial F}{\partial x}(\sigma_0,\, x_0)(-\tau(t)(x(t) - x_0)) \Big\| \\
&\le \sup_{\theta \in [0,\, 1]} \left\| \frac{\partial F}{\partial x}(\sigma(t),\, x(t) - \theta\tau(t)(x(t) - x_0)) - \frac{\partial F}{\partial x}(\sigma_0, x_0) \right\| \tau(t)\|x(t) - x_0\| \\
&\le L \left(\|td + \rho(t)\| + \sup_{\theta \in [0,\, 1]} \|x(t) - \theta\tau(t)(x(t) - x_0) - x_0\| \right) \tau(t)\|x(t) - x_0\| \\
&\le L \left(\|td + \rho(t)\| + \sup_{\theta \in [0,\, 1]} (1 - \theta\tau(t))\|x(t) - x_0\| \right) \tau(t)\|x(t) - x_0\| \\
&\le L \left(\left\| d + \frac{\rho(t)}{t} \right\| t + \|x(t) - x_0\| \right) \tau(t)\|x(t) - x_0\| \\
&\le L \left(\left\| d + \frac{\rho(t)}{t} \right\| t + \varepsilon_1 t^{1/2} \right) \varepsilon_1 \tau(t) t^{1/2} \\
&\le \left(L\varepsilon_1 \left\| d + \frac{\rho(t)}{t} \right\| t^{1/2} + L\varepsilon_1^2 \right) \tau(t) t \\
&\le \left(\frac{\delta}{6} + \frac{\delta}{6} \right) \tau(t) t \\
&= \frac{\delta}{3} \tau(t) t.
\end{aligned}$$

Similarly, by (4.1), (4.5), (4.10), the mean value theorem, and the definition of ε_1, we obtain

(4.14)
$$\begin{aligned}
\|\Phi_2(t)\| &\le \sup_{\theta \in [0,\, 1]} \|F'(\sigma_0 + \theta(td + \rho(t)),\, x_0 + \theta(x(t) - x_0)) \\
&\quad - F'(\sigma_0,\, x_0)\|(\|td + \rho(t)\| + \|x(t) - x_0\|) \\
&\le L \sup_{\theta \in [0,\, 1]} \theta(\|td + \rho(t)\| + \|x(t) - x_0\|)^2 \\
&\le L(\|td + \rho(t)\|^2 + 2\varepsilon_1 \|td + \rho(t)\| t^{1/2} + \varepsilon_1^2 t) \\
&= \left(L \left(\frac{\|td + \rho(t)\|^2}{t} + 2\varepsilon_1 \left\| d + \frac{\rho(t)}{t} \right\| t^{1/2} \right) + L\varepsilon_1^2 \right) t \\
&\le \left(\frac{\delta}{6} + \frac{\delta}{6} \right) t \\
&= \frac{\delta}{3} t.
\end{aligned}$$

We are now in a position to show that for $t \in (0,\, \bar{t}]$

$$F(\sigma(t),\, \tilde{x}(t)) - q(t) \in \operatorname{cone} B_\delta \left(\frac{\partial F}{\partial \sigma}(\sigma_0,\, x_0) d \right). \tag{4.15}$$

(This will mean that (2.13) is satisfied with $\sigma = \sigma(t)$, $x = \tilde{x}(t)$, and $\bar{y} = -\frac{\partial F}{\partial \sigma}(\sigma_0,\, x_0)d$.)

Indeed, put

$$y^1(t) = F(\sigma(t),\, x(t)) - p(t), \quad y^2(t) = \frac{\partial F}{\partial \sigma}(\sigma_0,\, x_0)\rho(t). \tag{4.16}$$

Then according to (4.12), (4.9), and (4.10) we have

$$\begin{aligned}
&F(\sigma(t),\, \tilde{x}(t))\\
&\qquad -q(t) = F(\sigma(t),\, \tilde{x}(t)) - F(\sigma(t),\, x(t)) + F(\sigma(t),\, x(t))\\
&\qquad\qquad + \tau(t)\frac{\partial F}{\partial x}(\sigma_0,\, x_0)(x(t) - x_0) - \tau(t)\frac{\partial F}{\partial x}(\sigma_0,\, x_0)(x(t) - x_0)\\
&\qquad\qquad - \tau(t)F(\sigma_0,\, x_0) - (1 - \tau(t))p(t)\\
&\qquad\qquad = \Phi_1(t) + F(\sigma(t),\, x(t)) - p(t)\\
&\qquad\qquad\quad - \tau(t)\left(\frac{\partial F}{\partial x}(\sigma_0,\, x_0)(x(t) - x_0) + F(\sigma_0,\, x_0) - p(t)\right)\\
&\qquad\qquad = \Phi_1(t) + y^1(t)\\
&\qquad\qquad\quad + \tau(t)\left(\Phi_2(t) - F(\sigma(t),\, x(t)) + t\frac{\partial F}{\partial \sigma}(\sigma_0,\, x_0)d + y^2(t) + p(t)\right)\\
&\qquad\qquad = \tau(t)t\left(\frac{\Phi_1(t)}{\tau(t)t} + \frac{y^1(t)}{\tau(t)t} + \frac{\Phi_2(t)}{t} - \frac{y^1(t)}{t} + \frac{y^2(t)}{t} + \frac{\partial F}{\partial \sigma}(\sigma_0,\, x_0)d\right)\\
&\qquad\qquad = \tau(t)t\left(\frac{\Phi_1(t)}{\tau(t)t} + (1 - \tau(t))\frac{y^1(t)}{\tau(t)t} + \frac{\Phi_2(t)}{t} + \frac{y^2(t)}{t} + \frac{\partial F}{\partial \sigma}(\sigma_0,\, x_0)d\right),
\end{aligned}$$

and hence, by (4.6), (4.7), (4.11), (4.13), (4.14), and (4.16),

$$\left\|\frac{F(\sigma(t),\, \tilde{x}(t)) - q(t)}{\tau(t)t} - \frac{\partial F}{\partial \sigma}(\sigma_0,\, x_0)d\right\| \le \frac{\delta}{3} + \frac{2\delta\,\mathrm{dist}(F(\sigma(t),\, x(t)),\, Q)}{12\,\mathrm{dist}(F(\sigma(t),\, x(t)))} + \frac{\delta}{3} + \frac{\delta}{6} = \delta,$$

which proves (4.15).

By the choice of $\varepsilon > 0$, $\delta > 0$, and $c > 0$, the estimate (1.3) holds with $\sigma = \sigma(t)$ and $x = \tilde{x}(t)$. Thus, taking into account (4.7), (4.8), we conclude that for all $t \in (0,\, \bar{t}]$

$$\begin{aligned}
\mathrm{dist}(x(t),\, D(\sigma(t))) &\le \|x(t) - \tilde{x}(t)\| + \mathrm{dist}(\tilde{x}(t),\, D(\sigma(t)))\\
&\le \|x(t) - \tilde{x}(t)\| + c\,\mathrm{dist}(F(\sigma(t),\, \tilde{x}(t)),\, Q)\\
&\le \|x(t) - \tilde{x}(t)\|\\
&\quad + c(\|F(\sigma(t),\, x(t)) - F(\sigma(t),\, \tilde{x}(t))\| + \mathrm{dist}(F(\sigma(t),\, x(t)),\, Q))\\
&\le \|x(t) - \tilde{x}(t)\| + c(l\|x(t) - \tilde{x}(t)\| + \mathrm{dist}(F(\sigma(t),\, x(t)),\, Q))\\
&\le (1 + c\,l)\|x(t) - x_0\|\tau(t) + c\,\mathrm{dist}(F(\sigma(t),\, x(t)))\\
&\le \left(c + \frac{12(1 + c\,l)}{\delta}\,\frac{\|x(t) - x_0\|}{t}\right)\mathrm{dist}(F(\sigma(t),\, x(t))).
\end{aligned}$$

This implies (4.3) with $a = \max\{c,\, 12(1 + c\,l)/\delta\}$. □

Note that the proof above actually specifies all the constants appearing in the assertion of Lemma 4.1.

The second result is [3, Lemma 4.109]. Our proof is an evident modification of the proof in [1, Lemma 6.2].

LEMMA 4.2. *Let Q be closed and convex, and let $x_0 \in D(\sigma_0)$. Let F be Fréchet-differentiable at $(\sigma_0,\, x_0)$ and Fréchet-differentiable with respect to x near $(\sigma_0,\, x_0)$, and let its derivative with respect to x be continuous at $(\sigma_0,\, x_0)$.*

If (1.6) *holds at* x_0 *with respect to a direction* $d \in \Sigma$*, then there exists* $a > 0$ *possessing the following property: for any mappings* $\rho(\cdot) : \mathbf{R}_+ \to \Sigma$ *and* $x(\cdot) : \mathbf{R}_+ \to X$ *such that* $\rho(t) = o(t)$*,* $x(t) \to x_0$ *as* $t \to 0$*, and the estimate*

$$\operatorname{dist}(F(\sigma_0 + td + \rho(t),\, x(t)),\, Q) = o(t) \tag{4.17}$$

holds for $t \geq 0$*, and for any* $\theta > 0$*, the estimate*

$$\operatorname{dist}(x(t),\, D(\sigma_0 + (1+\theta)td + \rho((1+\theta)t))) \leq a\theta t \tag{4.18}$$

holds for all $t > 0$ *small enough.*

Proof. By the same argument as in the proof of Lemma 4.1 we can choose $\varepsilon > 0$, $\delta > 0$, and $c > 0$ such that the estimate (1.3) holds for all $(\sigma,\, x) \in B_\varepsilon(\sigma_0) \times B_\varepsilon(\sigma_0)$ satisfying inclusion (2.13) with $\bar{y} = -\frac{\partial F}{\partial \sigma}(\sigma_0,\, x_0)d$. For each $t > 0$ put $\sigma(t) = \sigma_0 + td + \rho(t)$.

For a fixed $\theta > 0$ and for $t \geq 0$ we have

$$F(\sigma((1+\theta)t),\, x(t)) = F(\sigma(t),\, x(t)) + \theta t \frac{\partial F}{\partial \sigma}(\sigma_0,\, x_0)d + o(t). \tag{4.19}$$

Select $q(t) \in Q$ such that

$$\|F(\sigma(t),\, x(t)) - q(t)\| = \operatorname{dist}(F(\sigma(t),\, x(t)),\, Q) + o(t).$$

Then for $t > 0$ small enough $\sigma(t) \in B_\varepsilon(\sigma_0)$, $x(t) \in B_\varepsilon(x_0)$, and by (4.17) and (4.19) it holds that

$$\begin{aligned} F(\sigma((1+\theta)t),\, x(t)) - q(t) &= \theta t \frac{\partial F}{\partial \sigma}(\sigma_0,\, x_0)d + o(t) \\ &\in \operatorname{cone} B_\delta\left(\frac{\partial F}{\partial \sigma}(\sigma_0,\, x_0)d\right); \end{aligned} \tag{4.20}$$

i.e., inclusion (2.13) holds with $\sigma = \sigma((1+\theta)t)$, $x = x(t)$, and with $\bar{y} = -\frac{\partial F}{\partial \sigma}(\sigma_0,\, x_0)d$. Hence by (1.3) and the equality in (4.20) we conclude that

$$\begin{aligned} \operatorname{dist}(x(t),\, D(\sigma((1+\theta)t))) &\leq c \operatorname{dist}(F(\sigma((1+\theta)t),\, x(t))Q) \\ &\leq c\|F(\sigma((1+\theta)t),\, x(t)) - q(t)\| \\ &= c\left\|\frac{\partial F}{\partial \sigma}(\sigma_0,\, x_0)d\right\|\theta t + o(t), \end{aligned}$$

and the needed estimate (4.18) holds with any $a > c\|\frac{\partial F}{\partial \sigma}(\sigma_0,\, x_0)d\|$ for all $t > 0$ small enough. ∎

Acknowledgments. We would like to thank the anonymous referees for their helpful comments. Proposition 2.2 was entirely motivated by the referee's suggestion to try to remove an extra localization in condition (2.13) of Theorem 2.3 in the original version of the paper. The current version of Theorem 2.3 is of course equivalent to the original one, but its statement and the subsequent presentation became much more elegant. We are especially thankful to the referee for this.

REFERENCES

[1] A. V. Arutyunov and A. F. Izmailov, *Directional stability theorem and directional metric regularity*, Math. Oper. Res., 31 (2006), pp. 526–543.

[2] J. F. Bonnans and R. Cominetti, *Perturbed optimization in Banach spaces, Part* I: *A general theory based on a weak directional constraint qualification*, SIAM J. Control Optim., 34 (1996), pp. 1151–1171.

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

[4] J. M. Borwein, *Stability and regular points of inequality systems*, J. Optim. Theory Appl., 48 (1986), pp. 9–52.

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

[6] A. V. Dmitruk, A. A. Milyutin, and N. P. Osmolovskii, *Lyusternik's theorem and the theory of extrema*, Russian Math. Surveys, 35 (1980), pp. 11–51.

[7] A. L. Dontchev, *The Graves theorem revisited*, J. Convex Anal., 3 (1996), pp. 45–54.

[8] A. L. Dontchev, M. Quincampoix, and N. Zlateva, *Aubin criterion for metric regularity*, J. Convex Anal., 13 (2006), pp. 281–297.

[9] A. L. Dontchev and W. W. Hager, *Implicit functions, Lipschitz maps, and stability in optimization*, Math. Oper. Res., 19 (1994), pp. 753–768.

[10] P. C. Duong and H. Tuy, *Stability, surjectivity, and local invertibility of nondifferentiable mappings*, Acta Math. Vietnam., 3 (1978), pp. 89–105.

[11] B. Gollan, *On the marginal function in nonlinear programming*, Math. Oper. Res., 9 (1984), pp. 208–221.

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

[13] A. D. Ioffe, *Metric regularity and subdifferential calculus*, Russian Math. Surveys, 55 (2000), pp. 103–162.

[14] A. D. Ioffe and V. M. Tikhomirov, *Theory of Extremal Problems*, North–Holland, Amsterdam, The Netherlands, 1979.

[15] L. Lyusternik, *Conditional extrema of functions*, Math. USSR–Sb., 31 (1934), pp. 390–401.

[16] B. S. Mordukhovich, *Variational Analysis and Generalized Differentiation*, Springer-Verlag, Berlin, 2005.

[17] B. S. Mordukhovich, *Complete characterization of openness, metric regularity, and Lipschitzian properties of multifunctions*, Trans. Amer. Math. Soc., 340 (1993), pp. 1–36.

[18] B. S. Mordukhovich and B. Wang, *Restrictive metric regularity and generalized differential calculus in Banach spaces*, Int. J. Math. Math. Sci., 50 (2004), pp. 2650–2683.

[19] J.-P. Penot, *Metric regularity, openness, and Lipschitzian behavior of multifunctions*, Nonlinear Anal., 13 (1989), pp. 629–643.

[20] S. M. Robinson, *Regularity and stability of convex multivalued functions*, Math. Oper. Res., 1 (1976), pp. 130–143.

[21] S. M. Robinson, *Stability theory for systems of inequalities, Part* II: *Differentiable nonlinear systems*, SIAM J. Numer. Anal., 13 (1976), pp. 497–513.

[22] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.

[23] C. Ursescu, *Multifunctions with convex closed graph*, Czechoslovak Math. J., 25 (1975), pp. 438–441.

SIAM J. OPTIM.
Vol. 18, No. 3, pp. 834–852

BOUNDARY HALF-STRIPS AND THE STRONG CHIP*

EMIL ERNST† AND MICHEL THÉRA‡

Abstract. When the subdifferential sum rule formula holds for the indicator functions ι_C and ι_D of two closed convex sets C and D of a locally convex space X, the pair (C, D) is said to have the strong conical hull intersection property (the strong CHIP). The specification of a well-known theorem due to Moreau to the case of the support functionals σ_C and σ_D subsumes the fact that the pair (C, D) has the strong CHIP whenever the inf-convolution of σ_C and σ_D is exact. In this article we prove, in the setting of Euclidean spaces, that if the pair (C, D) has the strong CHIP while the boundary of C does not contain any half-strip, then the inf-convolution of σ_C and σ_D is exact. Moreover, when the boundary of a closed and convex set C does contain a half-strip, it is possible to find a closed and convex set D such that the pair (C, D) has the strong CHIP while the inf-convolution of σ_C and σ_D is not exact. The validity of the converse of Moreau's theorem in Euclidean spaces is thus associated with the absence of half-strips within the boundary of concerned convex sets.

Key words. strong conical hull intersection property, convex programming with convex inequalities, Euclidean space, exact infimal convolution, qualification conditions

AMS subject classifications. 90C46, 90C51, 46N10, 49K40

DOI. 10.1137/060658047

1. Introduction.

This study concerns an application of a geometrical notion called the *strong conical hull intersection property* (strong CHIP) introduced by Deutsch, Li, and Swetits (see [6]). We say that the pair (C, D) of closed and convex subsets of some locally convex space X has the strong CHIP if the subdifferential of the sum and sum of the subdifferentials of their indicator functions coincide:

$$\partial(\iota_C + \iota_D) = \partial\iota_C + \partial\iota_D. \tag{1.1}$$

As customary, ι_A is the indicator function of a subset A of X and is defined by $\iota_A(x) = 0$ if $x \in A$, and $\iota_A(x) = +\infty$ otherwise. We also recall that the convex subdifferential is an operator from X into the topological dual $X^\star$ of X, which assigns to each extended-real-valued mapping Φ on X a set-valued operator between X and $X^\star$ defined by

$$\partial\Phi(x_0) = \{x^\star \in X^* : \langle x - x_0, x^\star\rangle + \Phi(x_0) \le \Phi(x) \quad \forall x \in X\},$$

where $\langle\cdot,\cdot\rangle : X \times X^\star \to R$ is the duality pairing between X and $X^\star$.

Recalling that the normal cone $N_C(x)$ at $x \in X$ to a closed convex set C of X is the set $\partial\iota_C(x)$, the strong CHIP for the pair (C, D) amounts to saying that

$$N_{C\cap D}(x) = N_C(x) + N_D(x) \quad \forall x \in C \cap D; \tag{1.2}$$

i.e., every normal direction to $C \cap D$ at some point x can be expressed as the sum of normal directions at x to C and D.

*Received by the editors April 24, 2006; accepted for publication (in revised form) January 16, 2007; published electronically October 4, 2007.
http://www.siam.org/journals/siopt/18-3/65804.html
†Aix-Marseille Université, EA2596, Marseille, F-13397, France (Emil.Ernst@univ.u-3mrs.fr).
‡XLIM, Université de Limoges, 123 Avenue A. Thomas, 87060 Limoges Cedex, France (michel.thera@unilim.fr). The research of this author was partially supported by Agence Nationale de la Recherche under grant ANR NT05-1/43040.

This property is important in convex optimization because when we consider the problem of minimizing a convex functional Φ on the intersection of two sets C and D which have the strong CHIP, the optimality condition for $\bar{x}$ to be a minimizer becomes

$$0 \in \partial\Phi(\bar{x}) + N_C(\bar{x}) + N_D(\bar{x}).$$

In the case of convex differentiable optimization, it becomes

$$-\nabla f(\bar{x}) \in N_C(\bar{x}) + N_D(\bar{x}).$$

Let us quote only the result proved when X is a Hilbert space by Deutsch (see [4]). It says that the strong CHIP is the weakest constraint qualification under which a minimizer $\bar{x}$ of a convex function $\Phi : C_1 \cap C_2 \to R$ can be characterized using the subdifferential of Φ at x and the normal cones of C_1 and C_2 at x.

The existence of conditions ensuring that a pair of closed and convex sets has the strong CHIP is based on a classical result by Moreau ([9, Remarque 10.2]); the result initially published in [8] makes use of another key concept of convex analysis, namely the notion of *infimal convolution* (inf-convolution). Recall that, if Φ and Ψ are extended-real-valued lower semicontinuous convex functions over X (this class of functions from now on is denoted by $\Gamma_0(X)$), the inf-convolution of Φ and Ψ is the extended real-valued function $\Phi\square\Psi$ defined by

$$\Phi\square\Psi(x) = \inf_{y\in X} \left(\Phi(x-y) + \Psi(y)\right). \tag{1.3}$$

The infimal convolution between Φ and Ψ is said to be *exact* if $\Phi\square\Psi \in \Gamma_0(X)$ and the infimum is achieved in (1.3) whenever $\Phi\square\Psi(x) < +\infty$.

Let us also recall that given a convex closed set A in X, we note $\sigma_A : X^\star \to R \cup \{+\infty\}$, the support function of A. It is defined by

$$\sigma_A(f) = \sup_{x\in A} \langle x, f\rangle .$$

Using this concept, Moreau's theorem states that the subdifferential sum formula (1.1) holds, provided that the inf-convolution of the support functionals of C and D is exact. Remark that an equivalent way of expressing the exactness of the inf-convolution of the support functionals σ_C and σ_D is to say that every linear functional $f \in X^\star$ which is bounded above on $C \cap D$ may be expressed as the sum of two linear functionals f_1 and f_2, bounded above on C and D, respectively, such that

$$\sup_{x\in C\cap D} \langle x, f\rangle = \sup_{x\in C} \langle x, f_1\rangle + \sup_{x\in D} \langle x, f_2\rangle .$$

Let us observe that relation (1.2), i.e., the strong CHIP, is equivalent to the following property: *Every linear functional $f \in X^\star$ which achieves its maximum on $C \cap D$ may be expressed as the sum of two linear functionals f_1 and f_2, achieving their maximum on C and D, respectively, such that*

$$\max_{x\in(C\cap D)} \langle x, f\rangle = \max_{x\in C} \langle x, f_1\rangle + \max_{x\in D} \langle x, f_2\rangle .$$

The importance of Moreau's theorem comes from the fact that several very general *qualification conditions* are known to ensure the exactness of the inf-convolution of support functionals (the reader is referred for further information to the excellent

articles of Zălinescu [13], Gowda and Teboulle [10], and, respectively, Simons [11], in which he/she may find a clear picture of the topic, as well as self-contained proofs for most of the concerned results).

Accordingly, the result proved by Moreau gives the possibility of systematically specifying every qualification condition as a criterion for the strong CHIP (see for instance [5, Proposition 2.3]).

Let us remark that the exactness of the inf-convolution of the support functionals is stronger than the simple strong CHIP. Indeed, Moreau's condition concerns all the continuous linear functionals bounded above on the intersection $C \cap D$, while the strong CHIP is formulated only in terms of those elements from $X^\star$ which achieve their maximum on $C \cap D$. The question is thus raised of the validity of the converse to this theorem.

The converse to Moreau's theorem obviously holds for sets C and D such that every linear and continuous map bounded above on any one of the sets $C \cap D$, C, or D necessarily achieve their maximums on this set. On this ground, a first partial converse of the Moreau result has recently been proved by Bauschke, Borwein, and Li for Hilbert spaces (see [1, Proposition 6.4]); the result was extended to the setting of Banach spaces by Burachik and Jeyakumar [3, Proposition 4.2]. Their result states that, if C and D is a pair of closed and convex *cones* with the strong CHIP, then the inf-convolution of their support functionals is always exact.

However, it is not necessary to impose to every linear and continuous map which is bounded above on any one of the sets $C \cap D$, C, or D, to achieve its maximum on this set, in order to ensure the validity of the converse to the above mentioned Moreau's theorem. It is the aim of this article to clearly define the best conditions under which the converse of Moreau's theorem holds. More precisely, we characterize all the closed and convex subsets C of an Euclidean space X such that the following converse of Moreau's theorem holds: If, for some closed and convex set D, the pair (C, D) has the strong CHIP, then the inf-convolution of σ_C and σ_D is exact.

Our main result states that the validity of the converse of Moreau's theorem is ensured if and only if the boundary of the closed and convex subset C of the Euclidean space X does not contain any half-strip (by *half-strip* we mean, as customary, the convex hull of two disjoint and parallel half-lines). Note that the class of closed and convex sets without boundary half-strips is rather large, as it contains—the list is not exhaustive—all the bounded sets, the strictly convex sets, or even the continuous sets in the sense of Gale and Klee—sets such that their support functional is continuous except at the origin (see [7]).

The outline of the paper is as follows. The case of closed and convex sets without boundary half-strips is considered in section 2. We prove (Theorem 2.3) that, if the pair (C, D) has the strong CHIP, and if the boundary of one of the sets, say C, does not contain any half-strip, then the inf-convolution of the support functions of C and D must be exact.

The last section is concerned with convex sets which do admit at least one boundary half-strip. If the boundary of a closed and convex set C contains some half-strip, then we give a construction of a closed and convex set D such that the pair (C, D) has the strong CHIP, while the inf-convolution of σ_C and σ_D fails to be exact.

2. Convex sets without boundary half-strips. Now let us first collect some conditions ensuring in every reflexive Banach space the validity of the converse of Moreau's theorem.

PROPOSITION 2.1. *Let C and D be a pair of closed and convex subsets of a reflexive Banach space X. If C and D have the strong CHIP, and at least one of the following conditions holds:*

(i) *$C \cap D$ is bounded;*

(ii) *$C \cap D$ is a flat;*

(iii) *$C \cap D$ is a half-line,*

then the inf-*convolution of the support functions of C and D is exact. In other words, the converse of Moreau's theorem is valid.*

Proof of Proposition 2.1. We need the following standard convex analysis result.

LEMMA 2.2. *Let C and D be two closed and convex subsets of a locally convex space X, and consider an element y of $X^\star$ expressed as the sum $y = y_1 + y_2$ of two normal vectors $y_1 \in N_C(x)$ and $y_2 \in N_D(x)$, for some $x \in C \cap D$. Then the* inf-*convolution of the support functionals is exact at y, that is,*

$$\sigma_C \square \sigma_D(y) = \sigma_C(y_1) + \sigma_D(y_2) = \langle x, y \rangle . \tag{2.1}$$

Proof of Lemma 2.2. As $y_1 \in \partial \iota_C(x)$ and $y_2 \in \partial \iota_D(x)$, relation $y = y_1 + y_2$ implies that $y \in \partial \iota_{C \cap D}(x)$. Thus

$$\sigma_C(y_1) = \langle x, y_1 \rangle , \ \sigma_2(y_2) = \langle x, y_2 \rangle , \ \sigma_{C \cap D}(y) = \langle x, y \rangle ,$$

and hence

$$\sigma_{C \cap D}(y) = \sigma_C(y_1) + \sigma_D(y_2). \tag{2.2}$$

Recall that $\sigma_{C \cap D} \leq \sigma_C \square \sigma_D$, which means that

$$\sigma_{C \cap D}(y) \leq \sigma_C \square \sigma_D(y). \tag{2.3}$$

Finally, use the definition of the inf-convolution and the fact that $y = y_1 + y_2$ to deduce that

$$\sigma_C \square \sigma_D(y) \leq \sigma_C(y_1) + \sigma_D(y_2). \tag{2.4}$$

Relation (2.1) follows from relations (2.2), (2.3), and (2.4). □

Let us now return to the proof of Proposition 2.1 and consider that case (i) holds; i.e., we suppose that the pair (C, D) has the strong CHIP and $C \cap D$ is bounded.

As, in addition, X is a reflexive Banach space, it is easy to see that, for every $y \in X^\star$, there is an $x \in C \cap D$ such that $y \in \partial \iota_{C \cap D}(x)$. The pair (C, D) has the strong CHIP, and thus $y = y_1 + y_2$, for some $y_1 \in \partial \iota_C(x)$ and $y_2 \in \partial \iota_D(x)$; we may therefore apply Lemma 2.2 and deduce that

$$\sigma_C \square \sigma_D(y) = \sigma_C(y_1) + \sigma_D(y_2) = \langle x, y \rangle . \tag{2.5}$$

On one hand, from relation (2.5) we observe that the $\Gamma_0(X^\star)$-functional $\sigma_C \square \sigma_D$ is real-valued on $X^\star$ and thus, as $X^\star$ is a reflexive Banach space, it follows that $\sigma_C \square \sigma_D$ is continuous. Taking into account that relation (2.5) implies, on the other hand, that the infimum is always attained in the expression of the inf-convolution, we conclude that the inf-convolution of the support functions σ_C and σ_D is exact.

Case (ii). Let L be the closed subspace of X parallel to the flat $C \cap D$ (that is, $C \cap D = x_0 + L$ for every $x_0 \in C \cap D$), and factorize X with respect to L.

The quotient space X/L, say $\hat{X}$, is again a reflexive Banach space. Since $x_0 + L \subset C$ for every $x_0 \in C$, and $x_0 + L \subset D$ for every $x_0 \in D$, it follows that $\hat{C}$ and $\hat{D}$, the

quotients of the sets C and D, are closed and convex subsets of $\hat{X}$; moreover, it is straightforward to prove that the pair $(\hat{C}, \hat{D})$ has the strong CHIP if and only if the same holds for the pair (C, D), and that the inf-convolution of the support functions of $\hat{C}$ and $\hat{D}$ is exact if and only if the inf-convolution of the support functions of C and D is exact.

But the intersection between $\hat{C}$ and $\hat{D}$ reduces to a singleton, and case (ii) is proved by applying the conclusion of case (i) to the pair $(\hat{C}, \hat{D})$.

Case (iii). Set $x_0 + R_+\overline{x}$ for the half-line $C \cap D$. Obviously, when $y \in X^\star$ and $\langle \overline{x}, y\rangle \leq 0$ we have $y \in \partial\iota_{C\cap D}(x_0)$. Use the fact that the pair (C, D) has the CHIP to deduce that $y = y_1 + y_2$ for some $y_1 \in \partial\iota_C(x_0)$ and $y_2 \in \partial\iota_D(x_0)$, together with Lemma 2.2, to infer that

$$\sigma_C \square \sigma_D(y) = \sigma_C(y_1) + \sigma_D(y_2) = \langle x_0, y\rangle \quad \forall y \in X^\star,\ \langle \overline{x}, y\rangle \leq 0. \tag{2.6}$$

In order to obtain a similar relation for the case $\langle \overline{x}, y\rangle > 0$, note that for every $z \in X^\star$ such that $\langle \overline{x}, z\rangle > 0$, it holds that $\sigma_C(z) = \sigma_D(z) = +\infty$. Moreover, the inequality $\langle \overline{x}, z + v\rangle > 0$ means that at least one of the inequalities $\langle \overline{x}, z\rangle > 0$ and $\langle \overline{x}, v\rangle > 0$ holds. Combining these two facts, we deduce that

$$\sigma_C(z) + \sigma_D(v) = +\infty \quad \forall z, v \in X^\star \text{ such that } \langle \overline{x}, z + v\rangle > 0,$$

which means that

$$\sigma_C \square \sigma_D(y) = +\infty \quad \forall y \in X^\star,\ \langle \overline{x}, y\rangle > 0. \tag{2.7}$$

Combining relations (2.6) and (2.7) yields

$$\sigma_C \square \sigma_D = \iota_{\{y \in X^\star :\, \langle \overline{x}, y\rangle \leq 0\}} + \langle x_0, \cdot\rangle\,,$$

which means that the inf-convolution of σ_C and σ_D is the sum between the indicator function of a half-space and a linear and continuous functional, and clearly belongs to $\Gamma_0(X^\star)$. Use once more relation (2.6) to see that the infimum in the expression of the inf-convolution is achieved, and conclude that the inf-convolution of σ_C and σ_D is exact. □

Apparently, Proposition 2.1 lists three completely disparate conditions, each one being sufficient in its own way for the validity of the converse of Moreau's theorem. The geometric notion of a *half-strip*, that is, a convex hull of two parallel and disjoint half-lines, allows us to spot a common property of cases (i), (ii), and (iii) in Proposition 2.1.

THEOREM 2.3. *Let C and D two closed and convex subsets of the Euclidean space X, and assume that the boundary of the set C does not contain any half-strip. If the pair (C, D) has the strong CHIP, then the* inf*-convolution of σ_C and σ_D is exact (in other words, the converse of Moreau's theorem holds).*

Proof of Theorem 2.3. When the intersection $C \cap D$ meets the interior of C, we specify the well-known Moreau–Rockafellar internal point condition (see [9, Chap. 6, section 6.8]) to prove that the inf-convolution of the support functionals is exact.

If $C \cap D$ is a part of the boundary of C, use—as the boundary of C does not contain any half-strip—the obvious fact that the only closed and convex subsets of an Euclidean space which do not contain any half-strip are the bounded sets, the half-lines, and the lines, and completely prove Theorem 2.3 by making use of Proposition 2.1. □

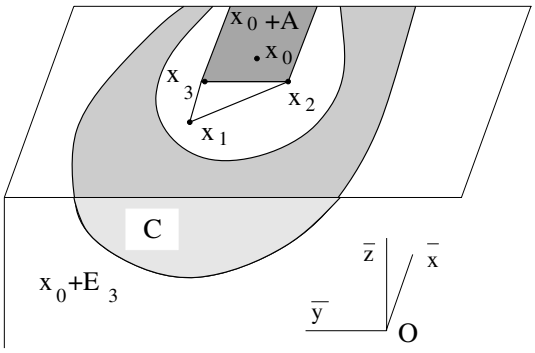


FIG. 3.1. *Closed and convex set with a boundary half-strip.*

3. Convex sets with boundary half-strips. The following result completes the analysis initiated in Theorem 2.3.

THEOREM 3.1. *Let C be a closed and convex subset of the Euclidean space X; assume moreover that the boundary of C contains at least one half-strip. Then there is a closed and convex set D such that the pair (C, D) has the CHIP while the* inf-*convolution between σ_C and σ_D is not exact.*

In other words, the existence of at least one boundary half-strip prevents the converse of Moreau's theorem from holding.

3.1. Construction and properties of the set D. Our strategy in this section is to construct the set D. The following easy result will be useful. It says that when X is Euclidean, every closed and convex set with a boundary half-strip may be contained within some half-space such that its boundary half-strip lies within the hyperplane which delimits this half-space.

PROPOSITION 3.2. *Let C be a closed and convex subset of an Euclidean space X such that its boundary contains a half-strip. Then, there is an orthonormal basis of X, $B = \{b_1, b_2, \ldots, b_n\}$, a positive parameter $a > 0$, and x_0, an element of X, such that*

$$x_0 + A \subset C \subset x_0 + E_3,$$

where A is the half-strip defined as

$$A = \{x \in X : 0 \leq x \cdot b_1,\ -2a \leq x \cdot b_2 \leq 2a,\ 0 = x \cdot b_i \quad \forall i \geq 3\},$$

and the half-space E_3 is given by the relation

$$E_3 = \{x \in X : \ x \cdot b_3 \leq 0\}.$$

Proof of Proposition 3.2. Let x_1, x_2, $\overline{x}$ in X be such that $\|\overline{x}\| = 1$ and the half-strip spanned by the half-lines $x_1 + R_+\overline{x}$ and $x_2 + R_+\overline{x}$ lies within the boundary of the set C. Assume (if necessary after changing x_1 into x_2) that $x_2 \cdot \overline{x} \geq x_1 \cdot \overline{x}$, and set $x_3 = x_1 + [(x_2 - x_1)\overline{x}]\overline{x}$.

Clearly, $x_3 \in x_1 + R_+\overline{x}$; as the half-lines $x_1 + R_+\overline{x}$ and $x_2 + R_+\overline{x}$ are disjoint, it follows that x_3 and x_2 cannot coincide. Set

$$a = \frac{\|x_3 - x_2\|}{4} \quad \text{and} \quad \overline{y} = \frac{x_3 - x_2}{4a} = \frac{x_3 - x_2}{\|x_3 - x_2\|}.$$

Note that

$$\overline{y}\cdot\overline{x}=\frac{1}{4a}(x_3\cdot\overline{x}-x_2\cdot\overline{x})=\frac{1}{4a}(x_1\cdot\overline{x}+(x_2-x_1)\overline{x}\,\overline{x}\cdot\overline{x}-x_2\cdot\overline{x})=0. \tag{3.1}$$

Finally, put

$$x_0=\frac{x_3+x_2}{2}+a\overline{x}=\frac{1}{2}(x_3+a\overline{x})+\frac{1}{2}(x_2+a\overline{x});$$

as $x_3+a\overline{x}\in x_1+R_+\overline{x}$ and $(x_2+a\overline{x})\in(x_2+R_+\overline{x})$, we deduce that x_0 belongs to the half-strip spanned by the half-lines $x_1+R_+\overline{x}$ and $x_2+R_+\overline{x}$, and thus belongs to the boundary of C.

Since x_0 is a boundary point of a closed and convex subset of an Euclidean space, it is well known that there is some linear mapping $\overline{z}\in X$, $\|\overline{z}\|=1$, which achieves its maximum on C at x_0,

$$\overline{z}\cdot x_0\geq\overline{z}\cdot x\qquad\forall x\in C. \tag{3.2}$$

Apply relation (3.2) for $x=x_0-a\overline{x}=\frac{x_3+x_2}{2}$ to deduce that $\overline{z}\cdot\overline{x}\geq 0$, and then for $x=x_0+a\overline{x}=\frac{1}{2}(x_3+2a\overline{x})+\frac{1}{2}(x_2+2a\overline{x})$ to obtain $\overline{z}\cdot\overline{x}\leq 0$ and therefore conclude that

$$\overline{z}\cdot\overline{x}=0. \tag{3.3}$$

Similarly, put $x_0-a\overline{x}-2a\overline{y}=x_2$ for x in relation (3.2), and using also relation (3.3), deduce that $\overline{z}\cdot\overline{y}\geq 0$. Finally, putting $x=x_0-a\overline{x}+2a\overline{y}=x_3$ in relation (3.2), and also taking into account relation (3.3), we infer $\overline{z}\cdot\overline{y}\leq 0$, that is,

$$\overline{z}\cdot\overline{y}=0. \tag{3.4}$$

Relations (3.1), (3.3), and (3.4) prove that it is possible to complete the set $\{\overline{x},\overline{y},\overline{z}\}$ depicted in Figure 3.1 up to $B=\{b_1,b_2,\ldots,b_n\}$, an orthonormal basis of X.

The proof of Proposition 3.2 will be completed if we remark that the set x_0+A is nothing but the half-strip spanned by the half-lines $(x_3+a\overline{x})+R_+\overline{x}$ and $(x_2+a\overline{x})+R_+\overline{x}$, and thus lies within the boundary of C, while relation (3.2) implies that C is a part of the half-space x_0+E_3. □

The basis B, the parameter a, and the element x_0 thus defined allow us to proceed to the construction of the set D. Let us first define the set F,

$$F=(P_1+S)\cap(P_2+T),$$

where $P_1\subset P_2$ are the sets bordered by two plane parabolae:

$$P_1=\left\{x\in X:\ x\cdot b_1\leq-\frac{a(x\cdot b_2)^2}{4},\ x\cdot b_i=0\quad\forall i\geq 3\right\},$$

$$P_2=\left\{x\in X:\ x\cdot b_1\leq-\frac{a(x\cdot b_2)^2}{8},\ x\cdot b_i=0\quad\forall i\geq 3\right\},$$

S is an orthogonal box in X:

$$S=\{x\in X:\ x\cdot b_1\leq 0,\ -1\leq ax\cdot b_2\leq 1,\ x\cdot b_3\leq 0\},$$

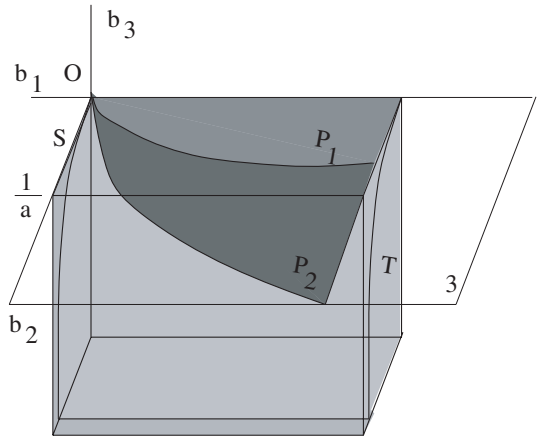


FIG. 3.2. *Sets needed in constructing the set D.*

T is a closed and convex subset of S:

$$T = \left\{ x \in X : \; x \cdot b_1 \le 0, \; -1 < ax \cdot b_2 < 1, \; x \cdot b_3 \le -\frac{a^2 (x \cdot b_2)^2}{1 - a^2 (x \cdot b_2)^2} \right\},$$

and, as customary, $Z^\star$ means the polar set of some subset Z of X,

$$Z^\star = \{ x \in X : \; x \cdot y \le 1 \quad \forall y \in Z \}.$$

Set now, as shown in Figure 3.2,

$$D = x_0 + F^\star = x_0 + ((P_1 + S) \cap (P_2 + T))^\star.$$

This definition grants to the set F (and thus D) several geometrical properties which are crucial for our purpose.

Let us first notice that F is contained in the half-space $E_1 = \{ x \in X : \; x \cdot b_1 \le 0 \}$; accordingly, the half-line $R_+ b_1$ lies within $F^\star$, and thus the half-line $x_0 + R_+ b_1$ is a part of both sets C and $D = x_0 + F^\star$. It follows that

$$\sigma_C(x) = \sigma_D(x) = +\infty \quad \forall x \in X \text{ such that } x \cdot b_1 > 0.$$

Let $x \in X$ be such that $x \cdot b_1 = 0$; if y is such that $y \cdot b_1 \ne 0$, then either $y \cdot b_1 > 0$ or $(x - y) \cdot b_1 > 0$, so

$$\sigma_C(y) + \sigma_D(x - y) = +\infty \quad \forall \, x \in X, \; x \cdot b_1 = 0, \; y \in X, \; y \cdot b_1 \ne 0; \tag{3.5}$$

hence for every element $x \in X$ such that $x \cdot b_1 = 0$ it results that

$$\sigma_C \square \sigma_D(x) = \inf_{y \in X, \; y \cdot b_1 = 0} \sigma_C(y) + \sigma_D(x - y). \tag{3.6}$$

The hyperplane $L_1 = \{ x \in X : \; x \cdot b_1 = 0 \}$ thus plays a very important role in computing the inf-convolution of the support functions σ_C and σ_D. The following lemma describes the intersection between the set F and L_1; for convenience, we state the result in terms of

$$\gamma_F(x) = \inf_{s > 0} \left\{ \frac{1}{s} : \; sx \in F \right\},$$

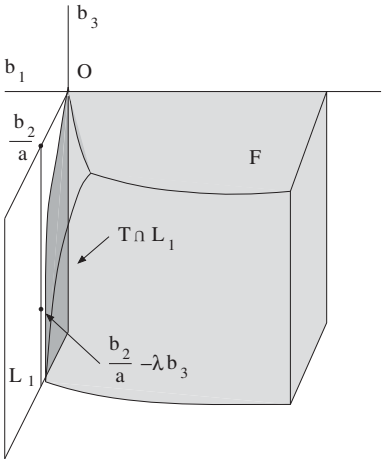


FIG. 3.3. *Intersection between F and L_1.*

the gauge function of the set F.

LEMMA 3.3. *The set F is a closed and convex subset of the Euclidean space X. Moreover, $F \cap L_1 = T \cap L_1$, and thus (see Figure* 3.3*)*

$$\gamma_F(x) \geq a|x \cdot b_2| \quad \forall x \in L_1, \tag{3.7}$$

$$\gamma_F(x) > 1 \quad \forall x \in L_1 \text{ such that } a|x \cdot b_2| = 1, \tag{3.8}$$

and

$$\lim_{\lambda \to +\infty} \gamma_F \left(\frac{b_2}{a} - \lambda b_3 \right) = 1. \tag{3.9}$$

Proof of Lemma 3.3. Recall that the sum $Z_1 + Z_2$ of two closed and convex subsets Z_1 and Z_2 of an Euclidean space is always convex. This sum is moreover closed, provided that Z_1 and $-Z_2$ do not contain two parallel half-lines (see [12, Corollary 9.1.2]). This is obviously the case for the pairs of closed and convex sets P_1 and S, as well as P_2 an T, and thus the sets $P_1 + S$ and $P_2 + T$ are closed and convex, and the same clearly holds also for the set F, which is their intersection.

It has already been noticed that all the sets P_1, P_2, S, and T lay within E_1; accordingly, the sum of two elements x_1 and x_2 from either P_1 and S, or P_2 and T, is contained within the delimiting hyperplane L_1 if and only if both elements x_1 and x_2 belong to L_1. In other words,

$$(P_1 + S) \cap L_1 = (P_1 \cap L_1) + (S \cap L_1),$$

$$(P_2 + T) \cap L_1 = (P_2 \cap L_1) + (T \cap L_1),$$

and note that $P_1 \cap L_1 = P_2 \cap L_1 = \{0\}$ to deduce that

$$\begin{aligned} F \cap L_1 &= ((P_1 + S) \cap L_1) \cap ((P_2 + T) \cap L_1) \\ &= (S \cap L_1) \cap (T \cap L_1) = (S \cap T) \cap L_1. \end{aligned}$$

Recall that $T \subset S$, and deduce that $F \cap L_1 = T \cap L_1$.

This relation may be used in order to compute the value of $\gamma_F(x)$ for elements $x \in L_1$, since it obviously holds that

$$\gamma_F(x) = \gamma_{F\cap L_1}(x) = \gamma_{T\cap L_1}(x) \quad \forall x \in L_1.$$

Use the fact that

$$T \cap L_1 \subset S \cap L_1 \subset M = \{x \in X : x \cdot b_1 = 0, -1 \le ax \cdot b_2 \le 1\}$$

to deduce that

$$\gamma_{T\cap L_1}(x) \ge \gamma_M(x) = a|x \cdot b_2| \quad \forall x \in L_1,$$

that is, relation (3.7).

In order to prove relation (3.8), note that, for every $x \in T \cap L_1$ we have $a|x\cdot b_2| < 1$. Accordingly, relation $a|x\cdot b_2| = 1$ implies that $x \notin T \cap L_1$. Let us now use [12, Corollary 9.7.1], which says that $T \cap L_1 = \{y : \ \gamma_{T\cap L_1}(y) \le 1\}$, and deduce that $\gamma_{T\cap L_1}(x) > 1$.

Finally, if $\lambda \ge 1$, standard computation shows

$$\frac{\sqrt{4\lambda^2+1}-1}{2\lambda}\left(\frac{b_2}{a} - \lambda b_3\right) \in T \cap L_1;$$

thus

$$\gamma_{T\cap L_1}\left(\frac{b_2}{a} - \lambda b_3\right) \le \frac{2\lambda}{\sqrt{4\lambda^2+1}-1}. \tag{3.10}$$

Use relation (3.8) for $x = \left(\frac{b_2}{a} - \lambda b_3\right)$ to see that

$$1 < \gamma_{T\cap L_1}\left(\frac{b_2}{a} - \lambda b_3\right); \tag{3.11}$$

relation (3.9) simply comes from relations (3.10) and (3.11). □

An important step in proving that the pair of closed and convex sets (C, D) has the strong CHIP is to determine their intersection $C \cap D$.

LEMMA 3.4. *It holds that*

$$C \cap D = x_0 + (F + Rb_3)^\star. \tag{3.12}$$

Proof of Lemma 3.4. Use the fact that $R_+(-b_3) \subset T \subset S$ to deduce that $R_+(-b_3) \subset F$, and thus that $F^\star \subset (-E_3)$. Accordingly, $D \subset x_0 + (-E_3)$, and as $C \subset x_0 + E_3$, we obtain that

$$C \cap D \subset (x_0 + E_3) \cap (x_0 + (-E_3)) = x_0 + L_3,$$

where by L_3 we mean

$$L_3 = \{x \in X : \ x \cdot b_3 = 0\}.$$

In other words,

$$x_0 \cdot b_3 = x \cdot b_3 \quad \forall x \in C \cap D. \tag{3.13}$$

Consequently,

$$C \cap D = C \cap (D \cap (x_0 + L_3)) = C \cap (x_0 + (F^\star \cap L_3)). \tag{3.14}$$

Recall (see [2, Chapter 4, section 1, Corollary of Proposition 3]) that, for every closed and convex sets A and B containing 0, it holds that $(A \cap B)^\star = \overline{\text{co}}(A^\star \cup B^\star)$, where $\overline{\text{co}}(A)$ denotes the closed convex hull of the set A.

Thus, by using the bipolar theorem (see [2, Chapter 4, section 1, Proposition 3]) applied to the set F, and the obvious fact that $L_3^\star = Rb_3$, we deduce that

$$(F^\star \cap L_3)^\star = \overline{\text{co}}(F^{\star\star} \cup L_3^\star) = \overline{\text{co}}(F \cup Rb_3). \tag{3.15}$$

It is well known that for every convex set A and flat W, $\overline{\text{co}}\,(A \cup W) = \overline{\text{co}}\,(A + W)$. Apply this relation to the convex set F and the one-dimensional flat Rb_3 to prove that $\overline{\text{co}}\,(F \cup Rb_3) = \overline{\text{co}}\,(F + Rb_3)$; by virtue of relation (3.15) it follows that

$$(F^\star \cap L_3)^\star = \overline{\text{co}}\,(F + Rb_3).$$

Accordingly, $F^\star \cap L_3 = (F^\star \cap L_3)^{\star\star} = (\overline{\text{co}}\,(F + Rb_3))^\star$; as the polar of any set coincides with the polar of its closure, we have

$$F^\star \cap L_3 = (F + Rb_3)^\star. \tag{3.16}$$

Let us prove that the set $(F + Rb_3)^\star$ lies within C. Indeed, after an easy computation it results that

$$\begin{aligned} \overline{(T + Rb_3)} &= S + Rb_3 \\ &= N = \{x \in X : \ x \cdot b_1 \le 0,\, -1 \le ax \cdot b_2 \le 1\}; \end{aligned} \tag{3.17}$$

thus $N \subset \overline{(F + Rb_3)}$, which means that $(F + Rb_3)^\star \subset N^\star$.

But as

$$N^\star = \{x \in X : \, 0 \le x \cdot b_1,\, -a \le x \cdot b_2 \le a,\, 0 = x \cdot b_i \quad \forall i \ge 3\},$$

we have (see Proposition 3.2) $N^\star \subset A$, and thus

$$x_0 + (F + Rb_3)^\star \subset (x_0 + N^\star) \subset (x_0 + A) \subset C. \tag{3.18}$$

Relation (3.12) follows now from relations (3.14), (3.16), and (3.18). □

It thus becomes necessary to determine the sum between the closed and convex set F and the line Rb_3.

LEMMA 3.5. *It holds that*

$$\{x \in (P_1 + S) : \, x \cdot b_1 < 0\} + Rb_3 \subset F + Rb_3 \subset (P_1 + S) + Rb_3; \tag{3.19}$$

accordingly,

$$C \cap D = x_0 + (P_1 + S + Rb_3)^\star. \tag{3.20}$$

Moreover, the gauge functions γ_F and $\gamma_{P_1+S+Rb_3}$ fulfill the following property: For every $x \in X$ such that $x \cdot b_1 < 0$, there is $\theta(x) \ge 0$ such that

$$\gamma_{P_1+S+Rb_3}(x) = \gamma_F\left(x - \theta(x) b_3\right). \tag{3.21}$$

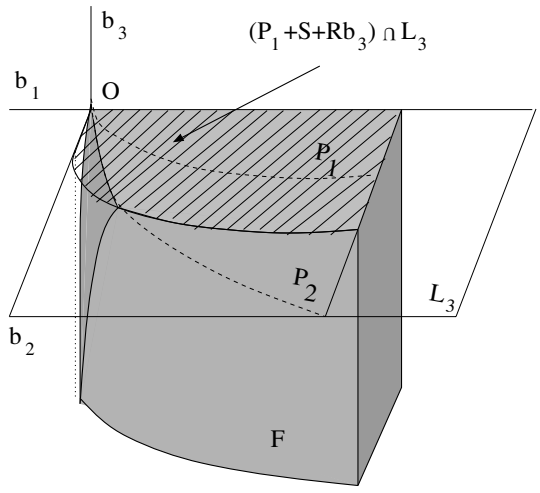


FIG. 3.4. *The sum between F and the line Rb_3.*

Proof of Lemma 3.5. On the basis of formula (3.17), we claim that $S + Rb_3$ is a closed and convex set. Moreover, there are no parallel half-lines within P_1 and $-(S + Rb_3)$, so, using again [12, Corollary 9.1.2], we deduce that the set $P_1 + S + Rb_3$ (see Figure 3.4) is closed and convex.

Let us prove the second inclusion in (3.19). As $F \subset (P_1 + S)$, it clearly follows that

$$F + Rb_3 \subset P_1 + S + Rb_3. \tag{3.22}$$

To establish the first inclusion in relation (3.19), we prove and use the fact that, for every $x \in P_1 + S$ such that $x \cdot b_1 < 0$, there is $\lambda(x) \geq 0$ such that $(x - \lambda(x)b_3) \in F$.

When $-1 < ax \cdot b_2 < 1$, it is easy to see that the value

$$\lambda(x) = x \cdot b_3 + \frac{a^2(x \cdot b_2)^2}{1 - a^2(x \cdot b_2)^2}$$

does the job. Indeed, the element $y = (x \cdot b_1)b_1$ lies within both P_1 and P_2, while

$$z = x - y - \lambda(x)b_3 = (x \cdot b_2)b_2 - \frac{a^2(x \cdot b_2)^2}{1 - a^2(x \cdot b_2)^2}b_3 + \sum_{i=4}^{n}(x \cdot b_i)b_i$$

is obviously contained in T, and thus in S. Accordingly,

$$x - \lambda(x)b_3 = y + (x - y - \lambda(x)b_3) \in (P_1 + S) \cap (P_2 + T) = F.$$

Let $x \in (P_1 + S)$ such that $x \cdot b_1 < 0$ and $a|x \cdot b_2| \geq 1$; to fix the ideas, admit that $ax \cdot b_2 \geq 1$. In order to define $\lambda(x)$ in this case, use the fact that x can be expressed as the sum $x = y + z$ of two elements y and z such that $y \in P_1$ and $z \in S$.

As $y \in P_1$, it follows that

$$y \cdot b_1 \leq -\frac{a(y \cdot b_2)^2}{4}; \tag{3.23}$$

since for every $z \in S$ it holds that $z \cdot b_1 \leq 0$, we deduce that thus $x \cdot b_1 \leq y \cdot b_1$. We may accordingly infer from relation (3.23) that

$$x \cdot b_1 \leq -\frac{a(y \cdot b_2)^2}{4}. \tag{3.24}$$

Use once more the fact that $z \in S$, to conclude that $-1 \leq az \cdot b_2 \leq 1$. Recall that $ax \cdot b_2 \geq 1$, and deduce that

$$(y \cdot b_2)^2 = (x \cdot b_2 - z \cdot b_2)^2 \geq \left(x \cdot b_2 - \frac{1}{a}\right)^2; \tag{3.25}$$

from relation (3.24) and (3.25) it follows that

$$x \cdot b_1 \leq -\frac{a\left(x \cdot b_2 - \frac{1}{a}\right)^2}{4}. \tag{3.26}$$

Combine the fact that $x \cdot b_1 \neq 0$ with relation (3.26) and deduce that

$$x \cdot b_1 < -\frac{a\left(x \cdot b_2 - \frac{1}{a}\right)^2}{8};$$

accordingly, for some parameter α such that $0 < a\alpha < 1$, we have

$$x \cdot b_1 < -\frac{a\left(x \cdot b_2 - \alpha\right)^2}{8}. \tag{3.27}$$

We can now define $\lambda(x)$ as

$$\lambda(x) = \frac{a^2\alpha^2}{1 - a^2\alpha^2}.$$

Inequality (3.27) proves that the element $(x - \alpha b_2 - \sum_{i=4}^{n}(x \cdot b_i)b_i)$ belongs to the set P_2; as obviously

$$\alpha b_2 - \frac{a^2\alpha^2}{1 - a^2\alpha^2}\, b_3 + \sum_{i=4}^{n}(x \cdot b_i)b_i \in T,$$

we deduce that

$$\begin{aligned} &x - \lambda(x)b_3 \\ &\quad = \left(x - \alpha b_2 - \sum_{i=4}^{n}(x \cdot b_i)b_i\right) + \left(\alpha b_2 - \frac{a^2\alpha^2}{1 - a^2\alpha^2}\, b_3 + \sum_{i=4}^{n}(x \cdot b_i)b_i\right) \\ &\quad \in P_2 + T. \end{aligned} \tag{3.28}$$

Remark that the case $ax \cdot b_2 \leq -1$ is similar to the case $ax \cdot b_2 \geq 1$. Indeed, when $ax \cdot b_2 \leq -1$, one has

$$(y \cdot b_2)^2 \geq \left(x \cdot b_2 + \frac{1}{a}\right)^2$$

instead of relation (3.25). The parameter α now lies between $-\frac{1}{a}$ and 0, and fulfills

$$x \cdot b_1 < -\frac{a\left(x \cdot b_2 + \alpha\right)^2}{8},$$

and not relation (3.26).

As in the case $ax \cdot b_2 \geq 1$, x is the sum of two elements: one in P_2, the other in T. However, when $ax \cdot b_2 \leq -1$, the element belonging to P_2 is $(x + \alpha b_2 - \sum_{i=4}^{n}(x \cdot b_i)b_i)$, and the one lying in T is $-\alpha b_2 - \frac{a^2\alpha^2}{1-a^2\alpha^2}\, b_3 + \sum_{i=4}^{n}(x \cdot b_i)b_i$.

Noticing that $S + R_+(-b_3) = S$, we have $(P_1 + S + R_+ b_3) = (P_1 + S)$. Thus, as $x \in P_1 + S$, we deduce that

$$x - \lambda b_3 \in P_1 + S \quad \forall \lambda \geq 0; \tag{3.29}$$

from (3.28) and (3.29) it follows that

$$(x - \lambda(x)b_3) \in (P_1 + S) \cap (P_2 + T) = F,$$

and therefore for every $x \in P_1 + S$ such that $x \cdot b_1 < 0$, there is $\lambda(x) \geq 0$ such that $x - \lambda(x)b_3 \in F$.

Use this observation to prove that

$$\{x \in P_1 + S : x \cdot b_1 < 0\} + Rb_3 \subset F + Rb_3,$$

which, together with relation (3.22), yields relation (3.19).

Relation (3.19) implies that the set $P_1 + S + Rb_3$ is the closure of the set $F + Rb_3$. Recalling that the polar of any set coincides with the polar of its closure, we deduce that

$$(F + Rb_3)^\star = (P_1 + S + Rb_3)^\star, \tag{3.30}$$

and relation (3.20) follows from formulas (3.12) and (3.30).

It remains to prove relation (3.21). To begin with, notice that from relation (3.19), it follows that $F \subset P_1 + S + Rb_3$, and thus

$$\gamma_{P_1+S+Rb_3} \leq \gamma_F. \tag{3.31}$$

Let us first prove that $\gamma_{P_1+S+Rb_3}(x)$ is real-valued for every $x \in X$ such that $x \cdot b_1 < 0$. In this respect, note that from relation (3.17) it results that $\{x \in X : x \cdot b_2 = x \cdot b_3 = 0\} \subset S + Rb_3$, and thus that

$$P_1 + \{x \in X : x \cdot b_1 = x \cdot b_2 = 0\} \subset (P_1 + S + Rb_3). \tag{3.32}$$

On the other hand, the set $P_1 + \{x \in X : x \cdot b_1 = x \cdot b_2 = 0\}$ contains all the elements $x \in X$ such that $x \cdot b_1 \leq -\frac{a(x \cdot b_2)^2}{4}$. As

$$\left(-\frac{4x \cdot b_1}{a(x \cdot b_2)^2}\, x\right) \cdot b_1 = -\frac{4(x \cdot b_1)^2}{a(x \cdot b_2)^2} = -\frac{a\left(\left(-\frac{4x \cdot b_1}{a(x \cdot b_2)^2}\, x\right) \cdot b_2\right)^2}{4},$$

this means that

$$-\frac{4x \cdot b_1}{a(x \cdot b_2)^2} x \in P_1 + \{x \in X : x \cdot b_1 = x \cdot b_2 = 0\}.$$

Combine the previous relation with formula (3.32) to deduce that

$$\gamma_{P_1+S+Rb_3}(x) \leq \frac{a(x \cdot b_2)^2}{|4x \cdot b_1|}.$$

Consequently, for every $x \in X$ such that $x \cdot b_1 < 0$ we have $\gamma_{P_1+S+Rb_3}(x) < +\infty$.

Let us first consider the case when $\gamma_{P_1+S+Rb_3}(x) = 0$, that is, when x belongs to a ray completely contained in $P_1 + S + Rb_3$. Taking into account the definitions of the sets P_1 and S, we have

$$[\gamma_{P_1+S+Rb_3}(x) = 0] \Leftrightarrow [x \cdot b_1 \leq 0 \text{ and } x \cdot b_2 = 0]\,;$$

similarly,

$$[\gamma_F(x) = 0] \Leftrightarrow [x \cdot b_1 \leq 0, x \cdot b_2 = 0 \text{ and } x \cdot b_3 \leq 0]\,.$$

In this case, $\theta(x) = \frac{x \cdot b_3 + |x \cdot b_3|}{2}$ obviously does the job.

Let us now turn to the case $\gamma_{P_1+S+Rb_3}(x) > 0$ and write that

$$\frac{x}{\gamma_{P_1+S+Rb_3}(x)} \in P_1 + S + Rb_3.$$

We deduce that there is a $\tilde{\lambda}(x) \in R$ such that

$$\frac{x}{\gamma_{P_1+S+Rb_3}(x)} - \tilde{\lambda}(x)b_3 \in P_1 + S.$$

Accordingly,

$$\left(\frac{x}{\gamma_{P_1+S+Rb_3}(x)} - \tilde{\lambda}(x)b_3\right) - \lambda\left(\frac{x}{\gamma_{P_1+S+Rb_3}(x)} - \tilde{\lambda}(x)b_3\right) b_3 \in F,$$

which means that

$$\gamma_F\left(x - \gamma_{P_1+S+Rb_3}(x)\left(\tilde{\lambda}(x) + \lambda\left(\frac{x}{\gamma_{P_1+S+Rb_3}(x)} - \tilde{\lambda}(x)b_3\right)\right) b_3\right) \leq \gamma_{P_1+S+Rb_3}(x). \tag{3.33}$$

Relations (3.31), (3.33) and the obvious fact that

$$\gamma_{P_1+S+Rb_3}(x) = \gamma_{P_1+S+Rb_3}(x + \nu b_3) \quad \forall \nu \in R$$

prove relation (3.21) with

$$\theta(x) = \gamma_{P_1+S+Rb_3}(x)\left(\tilde{\lambda}(x) + \lambda\left(\frac{x}{\gamma_{P_1+S+Rb_3}(x)} - \tilde{\lambda}(x)b_3\right)\right),$$

completing in this way the proof of Lemma 3.5. □

3.2. The main result. We claim that *the pair of closed and convex sets C and D has the strong CHIP.*

PROPOSITION 3.6. *The pair of closed and convex subsets C and D of the Euclidean space X has the strong CHIP.*

Proof of Proposition 3.6. Let $x_1 \in C \cap D$ and $y \in \partial\iota_{C\cap D}(x_1)$, $y \neq 0$. Our aim is to express y as the sum of two elements y_1 and y_2 from $\partial\iota_C(x_1)$ and $\partial\iota_D(x_1)$.

Let us first remark that, since $(x_0 + A) \subset C \subset (x_0 + E_3)$ (see Proposition 3.2), it follows that

$$R_+b_3 \subset \partial\iota_C(x) \tag{3.34}$$

for every $x \in x_0 + A$, in particular for every $x \in C \cap D$ (see Lemma 3.4 and relation (3.18)). Similarly, we deduce that

$$R_+(-b_3) \subset \partial\iota_D(x) \tag{3.35}$$

for every $x \in C \cap D$. The flat $\{x \in X :\ x \cdot b_i = 0\ \forall 1 \le i \le 3\}$ obviously lies within T, and thus in S and F. Thus D is contained within the flat $x_0 + \{x \in X :\ x \cdot b_i = 0\ \forall i \ge 4\}$, and we deduce that

$$\{x \in X :\ x \cdot b_1 = x \cdot b_2 = x \cdot b_3 = 0\} \subset \partial\iota_D(x) \quad \forall x \in D. \tag{3.36}$$

From relations (3.34), (3.35), and (3.36) it follows that

$$\{x \in X :\ x \cdot b_1 = x \cdot b_2 = 0\} \subset (\partial\iota_C(x) + \partial\iota_D(x)) \quad \forall x \in (C \cap D). \tag{3.37}$$

We address first the case when $y \cdot b_1 = 0$. A standard computation shows that

$$(P_1 + S + Rb_3)^\star \tag{3.38}$$

$$= \left\{ x \in X : x \cdot b_1 \ge \frac{(x \cdot b_2)^2}{a - |x \cdot b_2|}, -a < x \cdot b_2 < a,\ x \cdot b_i = 0 \quad \forall i \ge 3 \right\}.$$

From the previous relation it follows that the set $(P_1 + S + Rb_3)^\star$ is contained within the plane spanned by b_1 and b_2. For every $y \in X$ such that $y \cdot b_1 = 0$ it follows that

$$x \cdot y = (x \cdot b_2)(y \cdot b_2) \quad \forall x \in (P_1 + S + Rb_3)^\star. \tag{3.39}$$

The elements x_1 and x_0 are both in $C \cap D$; in view of relation (3.20) it appears that $(x_1 - x_0) \in (P_1 + S + Rb_3)^\star$.

From relation (3.38) it follows that $|(x_1 - x_0) \cdot b_2| < a$. Set

$$\alpha = \frac{a + |(x_1 - x_0) \cdot b_2|}{2},\ z_1 = \frac{\alpha^2}{a - \alpha} b_1 - \alpha b_2,\ z_2 = \frac{\alpha^2}{a - \alpha} b_1 + \alpha b_2;$$

thus $z_1, z_2 \in (P_1 + S + Rb_3)^\star$ and $z_1 \cdot b_2 < (x_1 - x_0) \cdot b_2 < z_2 \cdot b_2$.

Recall that, as $y \in \partial\iota_{C\cap D}(x_1)$, the linear functional $X \ni x \to (x \cdot y) \in R$ achieves its maximum on $C \cap D$ at x_1. Thus, on one hand,

$$(x_0 + z_1) \cdot y \le x_1 \cdot y,$$

that is, in view of relation (3.39),

$$(z_1 \cdot b_2)(y \cdot b_2) \le ((x_1 - x_0) \cdot b_2)\,(y \cdot b_2);$$

combine this relation with the fact that $z_1 \cdot b_2 < (x_1 - x_0) \cdot b_2$, and deduce that $y \cdot b_2 \ge 0$. On the other hand,

$$(x_0 + z_2) \cdot y \le x_1 \cdot y,$$

that is, once more by virtue of relation (3.39),

$$(z_2 \cdot b_2)(y \cdot b_2) \le ((x_1 - x_0) \cdot b_2)\,(y \cdot b_2);$$

in addition, as $(x_1 - x_0) \cdot b_2 < z_2 \cdot b_2$, we get $y \cdot b_2 \le 0$.

We may thus conclude that, when $y \cdot b_1 = 0$, it results that $y \cdot b_2 = 0$, and formula (3.37) proves that y is the sum of two elements from $\partial\iota_C(x_1)$ and $\partial\iota_D(x_1)$.

Consider now the case when $y \cdot b_1 \neq 0$, which, taking into account the fact that the half-line $(x_0 + R_+ b_1)$ is contained (as already remarked) within $C \cap D$, amounts to saying that $y \cdot b_1 < 0$.

It is well known (see [12, Theorem 14.5]) that for every closed and convex set Z containing 0, it holds $\sigma_Z = \gamma_{Z^\star}$. Use this relation for the set $(P_1 + S + Rb_3)^\star$ to obtain

$$\sigma_{(P_1+S+Rb_3)^\star} = \gamma_{P_1+S+Rb_3};$$

as (see 3.20) $(P_1 + S + Rb_3)^\star = (C \cap D) - x_0$, it follows that

$$\sigma_{(C\cap D)-x_0} = \gamma_{P_1+S+Rb_3}. \tag{3.40}$$

Similarly,

$$\sigma_{D-x_0} = \gamma_F. \tag{3.41}$$

From relations (3.21), (3.40), and (3.41) it follows that there is some $\theta(y) \geq 0$ such that

$$\sigma_{(C\cap D)-x_0}(y) = \sigma_{D-x_0}(y - \theta(y)b_3);$$

thus

$$\sigma_{C\cap D}(y) = \sigma_D(y - \theta(y)b_3) + \theta(y)x_0 \cdot b_3. \tag{3.42}$$

As $y \in \partial\iota_{C\cap D}(x_1)$, it results that $\sigma_{C\cap D}(y) = x_1 \cdot y$; use relation (3.42) to see that

$$\sigma_D(y - \theta(y)b_3) + \theta(y)x_0 \cdot b_3 = x_1 \cdot y. \tag{3.43}$$

Relation (3.13) reads $x_0 \cdot b_3 = x_1 \cdot b_3$. Equality (3.43) may thus be stated as

$$\sigma_D(y - \theta(y)b_3) + \theta(y)x_1 \cdot b_3 = x_1 \cdot y,$$

that is,

$$\sigma_D(y - \theta(y)b_3) = x_1 \cdot (y - \theta(y)b_3).$$

This means that $(y - \theta(y)b_3) \in \partial\iota_D(x_1)$.

Recall (see relation (3.34)) that $\lambda b_3 \in \partial\iota_C(x_1)$ for every $\lambda \geq 0$, and express y as $y = \theta(y)b_3 + (y - \theta(y)b_3)$, that is, the sum of an element from $\partial\iota_C(x_1)$ and the sum of another element from $\partial\iota_D(x_1)$. □

We finally claim that *the* inf-*convolution of the support functionals σ_C and σ_D is not exact at $\frac{b_2}{a}$*, a fact which completes the proof of Theorem 3.1.

PROPOSITION 3.7. *It holds that*

$$\sigma_C \square \sigma_D\left(\frac{b_2}{a}\right) = \frac{x_0 \cdot b_2}{a} + 1, \tag{3.44}$$

while

$$\sigma_C(y) + \sigma_D(z) > \frac{x_0 \cdot b_2}{a} + 1 \quad \forall y + z = \frac{b_2}{a}. \tag{3.45}$$

Proof of Proposition 3.7. Use the fact that (x_0+2ab_2) and (x_0-2ab_2) both belong to $x_0 + A$, and thus to C, to deduce that

$$\sigma_C(x) \geq \max\left((x_0 + 2ab_2) \cdot x, (x_0 - 2ab_2) \cdot x\right) = x_0 \cdot x + 2a|x \cdot b_2|. \tag{3.46}$$

From relation (3.7) it follows that

$$\gamma_F(x) \geq a|x \cdot b_2| \quad \forall x \in L_1.$$

Relation (3.41) reads $\sigma_{D-x_0} = \gamma_F$; hence, it results that

$$\sigma_D(x) = x_0 \cdot x + \gamma_F(x). \tag{3.47}$$

It follows that

$$\sigma_D(x) \geq x_0 \cdot x + a|x \cdot b_2| \quad \forall x \in L_1. \tag{3.48}$$

From relations (3.46) and (3.48) it results that

$$\begin{aligned} \sigma_C(x)+\sigma_D\left(\frac{b_2}{a} - x\right) & \\ \geq \; & x_0 \cdot x + 2a|x \cdot b_2| + x_0 \cdot \left(\frac{b_2}{a} - x\right) + a\left|\left(\frac{b_2}{a} - x\right) \cdot b_2\right| \\ \geq \; & \frac{x_0 \cdot b_2}{a} + 1 + a|x \cdot b_2| \quad \forall x \in X,\ x \cdot b_1 = 0. \end{aligned} \tag{3.49}$$

By taking into account relations (3.6) and (3.49) we prove that

$$\sigma_C \square \sigma_D\left(\frac{b_2}{a}\right) \geq \frac{x_0 \cdot b_2}{a} + 1. \tag{3.50}$$

As $x_0 \in C \subset (x_0 + E_3)$, it follows that

$$\sigma_C(\lambda b_3) = \lambda x_0 \cdot b_3 \quad \forall \lambda \geq 0.$$

Use relation (3.47) to deduce that

$$\begin{aligned} \sigma_C(\lambda b_3) + \sigma_D\left(\frac{b_2}{a} - \lambda b_3\right) & \\ = \lambda x_0 \cdot b_3 + \frac{x_0 \cdot b_2}{a} - \lambda x_0 \cdot b_3 + \gamma_F\left(\frac{b_2}{a} - \lambda b_3\right). & \end{aligned}$$

From the previous equality, together with relation (3.9), it yields that

$$\lim_{\lambda \to \infty}\left(\sigma_C(\lambda b_3) + \sigma_D\left(\frac{b_2}{a} - \lambda b_3\right)\right) = \frac{x_0 \cdot b_2}{a} + 1,$$

which, combined with inequality (3.50), proves relation (3.44).

Finally, let $x \in L_1$ be such that $x \cdot b_2 = 0$. Then (see relation (3.8))

$$\gamma_F\left(\frac{b_2}{a} - x\right) > 1,$$

and, as obviously $\sigma_C(x) \geq x_0 \cdot x$, it results that

$$\begin{aligned} \sigma_C(x) + \sigma_D\left(\frac{b_2}{a} - x\right) & \\ \geq x_0 \cdot x + \frac{x_0 \cdot b_2}{a} - x_0 \cdot x + \gamma_F\left(\frac{b_2}{a} - x\right) & \\ > \frac{x_0 \cdot b_2}{a} + 1 \quad \forall\, x \in X,\ x \cdot b_1 = x \cdot b_2 = 0. & \end{aligned} \tag{3.51}$$

Use relation (3.49) to deduce that, for every $x \in X$ such that $x \cdot b_1 = 0$ and $x \cdot b_2 \neq 0$, it holds that

$$\sigma_C(x) + \sigma_D\left(\frac{b_2}{a} - x\right) > \frac{x_0 \cdot b_2}{a} + 1. \tag{3.52}$$

Relation (3.45) follows from relations (3.5), (3.52), and (3.51). □

Acknowledgments. We would like to warmly thank the two anonymous referees. Their careful reading of the paper allowed us to correct a significant number of typos and errors and largely contributed to the final form of the article.

REFERENCES

[1] H. H. Bauschke, J. M. Borwein, and W. Li, *The strong conical hull intersection property, bounded linear regularity, Jameson's property (G), and error bounds in convex optimization*, Math. Program. Ser. A, 86 (1999), pp. 135–160.
[2] N. Bourbaki, *Eléments de mathématique*. XVIII, Actualités Sci. Ind. 1229, Hermann & Cie, Paris, 1955.
[3] R. S. Burachik and V. Jeyakumar, *A simple closure condition for the normal cone intersection formula*, Proc. Amer. Math. Soc., 133 (2005), pp. 1741–1748.
[4] F. Deutsch, *The role of the strong conical hull intersection property in convex optimization and approximation*, in Approximation Theory IX, Innov. Appl. Math., Vanderbilt University Press, Nashville, TN, 1998, pp. 105–112.
[5] F. Deutsch, W. Li, and J. D. Ward, *Best approximation from the intersection of a closed convex set and a polyhedron in Hilbert space, weak Slater conditions, and the strong conical hull intersection property*, SIAM J. Optim., 10 (1999), pp. 252–268.
[6] F. Deutsch, W. Li, and J. Swetits, *A dual approach to constrained interpolation from a convex subset of Hilbert space*, J. Approx. Theory, 90 (1997), pp. 385–414.
[7] D. Gale and V. Klee, *Continuous convex sets*, Math. Scand., 7 (1959), pp. 370–391.
[8] J. J. Moreau, *Étude locale d'une fonctionnelle convexe*, Université de Montpellier, Montpellier, France, 1963.
[9] J. J. Moreau, *Fonctionnelles convexes, séminaire sur les équations aux dérivées partielles*, Collège de France, Paris, 1967.
[10] M. S. Gowda and M. Teboulle, *A comparison of constraint qualifications in infinite-dimensional convex programming*, SIAM J. Control Optim., 28 (1990), pp. 925–935.
[11] S. Simons, *Sum theorems for monotone operators and convex functions*, Trans. Amer. Math. Soc., 350 (1998), pp. 2953–2972.
[12] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1968.
[13] C. Zălinescu, *A comparison of constraint qualifications in infinite-dimensional convex programming revisited*, J. Austral. Math. Soc. Ser. B, 40 (1999), pp. 353–378.

SIAM J. OPTIM.
Vol. 18, No. 3, pp. 853–877

AMBIGUOUS RISK MEASURES AND OPTIMAL ROBUST PORTFOLIOS*

GIUSEPPE C. CALAFIORE†

Abstract. This paper deals with a problem of guaranteed (robust) financial decision-making under model uncertainty. An efficient method is proposed for determining optimal robust portfolios of risky financial instruments in the presence of ambiguity (uncertainty) on the probabilistic model of the returns. Specifically, it is assumed that a *nominal* discrete return distribution is given, while the *true* distribution is only known to lie within a distance d from the nominal one, where the distance is measured according to the Kullback–Leibler divergence. The goal in this setting is to compute portfolios that are worst-case optimal in the mean-risk sense, that is, to determine portfolios that minimize the maximum with respect to all the allowable distributions of a weighted risk-mean objective. The analysis in the paper considers both the standard variance measure of risk and the absolute deviation measure.

Key words. worst-case financial risk, portfolio selection, asset allocation, statistical ambiguity, robust optimization

AMS subject classifications. 91B28, 90B50, 90C25, 90C47, 90C51, 90C90

DOI. 10.1137/060654803

1. Introduction.

A classical problem in computational finance is that of optimally selecting a portfolio of finitely many risky assets so as to maximize the expected return of the investment while keeping "risk" under control. In the mainstream approach, dating back to the seminal work of Markowitz [25], risk is measured according to the variance of the portfolio return, and the determination of an optimal portfolio amounts to the solution of a convex quadratic programming problem. Since the introduction of this basic mean-variance model for portfolio selection, however, many criticisms have been raised on its practical relevance, especially in regard to the sensitivity of the optimal portfolios with respect to the statistical errors in the parameters (the estimated expected returns and covariances of the assets), and possible remedies have been proposed. An in-depth overview of this literature is out of the scope of the present work, but the interested reader could find some useful pointers in [3, 6, 8, 26].

More recently, the issue of model uncertainty in portfolio optimization has been the subject of study from different groups of researchers; see, for instance, [13, 18, 24, 34, 37]. Many of these recent contributions propose ideas and computational tools derived from the robust convex optimization field [4, 14]. The general approach in this setting is to model the uncertain market parameters (expected returns and covariances) as deterministic unknown-but-bounded quantities, and then take a worst-case approach where an optimal portfolio is sought that minimizes the worst risk that the investor may face as the market parameters vary in any possible way inside their admissible domains. These deterministic models are practically and theoretically sound, since they either are naturally derived from confidence regions around the least-squares estimates of the market parameters (see [18]) or may reflect an analyst feeling

*Received by the editors March 22, 2006; accepted for publication (in revised form) February 5, 2007; published electronically October 4, 2007. This work was supported by MIUR under the FIRB project "Learning, Randomization and Guaranteed Predictive Inference for Complex Uncertain Systems."

http://www.siam.org/journals/siopt/18-3/65480.html

†Dipartimento di Automatica e Informatica, Politecnico di Torino, 10129 Torino, Italy (giuseppe.calafiore@polito.it).

of the reliability of the parameter estimates. In this latter case, the uncertainty model typically takes the form of elementwise bounds on some or all entries of the expected return vector and covariance matrix; see [13, 34]. Specifically, El Ghaoui, Oks, and Oustry in [13] consider the problem of computing and optimizing the worst-case value-at-risk of a portfolio, under bounded uncertainty on the mean and covariance matrix of the returns, and show how the computation can be efficiently performed by recasting the problem in the form of a semidefinite optimization program [33, 35]. Goldfarb and Iyengar in [18] develop a robust factor model for the returns, show how the uncertainty description can be naturally obtained from confidence regions of standard statistical estimation techniques, and pose the corresponding robust allocation problem in the form of a convex second order cone program (SOCP); see [23]. Tütüncü and Koenig in [34] propose the use of an interval uncertainty model for the return mean and covariance and solve the resulting worst-case Markowitz problem via an ad hoc saddle-point algorithm.

While the mentioned approaches are specific to portfolio selection problems, more general models dealing with uncertainty in the underlying probability measures have a long history and have been studied in different fields, such as economics, finance, and stochastic optimization; see, e.g., [7, 11, 12, 32, 36]. Uncertainty in the probabilistic model is usually referred to as *ambiguity* in the decision theory literature. The recent work from Erdoğan and Iyengar [15] discusses ambiguous chance-constrained problems and employs the Prohorov metric to describe the uncertainty "ball" of admissible distributions. As we shall see in section 2.1, in this paper we adopt a similar approach for describing the "ambiguity" set around a nominal distribution and employ the Kullback–Leibler divergence function as a distance measure among distributions. This distance measure has nice invariance and convexity properties, and the degree of ambiguity in this metric can be estimated from samples; see [20].

The main goal of this paper is to present an efficient computational framework for robust portfolio selection in the situation of asset returns described by an ambiguous discrete joint probability distribution. We consider two risk measures (see [2, 30]) given by composite objectives of the form $\rho(x,\pi) - \gamma\mu(x,\pi)$, where $\rho(x,\pi)$ is either the variance or the expected absolute deviation of the portfolio, $\mu(x,\pi)$ is the portfolio expected return, and γ is a nonnegative parameter. Here, x denotes the portfolio mix and π the discrete distribution of the returns (see section 2 for precise definitions and notation). The measure based on the expected absolute deviation is (for $\gamma \geq 2$) a *coherent* measure of risk in the sense of Artzner et al. [2]. The measure based on the variance is instead not coherent, since it violates a monotonicity condition; see [30]. However, the use of this latter measure is justified by both historical reasons and its wide popularity.

In the *nominal* case—i.e., when the probability distribution π is known and given—minimizing the above objectives is equivalent either to a standard Markowitz problem (in the case of the variance-based risk measure) or to the absolute deviation problem, discussed, for instance, in [21, 31]. It is well known that, in this latter case, the optimal portfolio can be found by solving a linear programming problem.

The key point in this paper is to consider the return distribution π to be imprecisely known. In particular, we assume that a nominal value η for the distribution is given, but that the actual π is only known to lie in a region at distance no larger than d from its nominal value, where d is a user-definable parameter that quantifies the (lack of) confidence in the nominal probability (the "index of ambiguity," in the terminology of [20]). To measure the distance among distributions, we use the stan-

dard metric given by the Kullback–Leibler divergence. In this setting, we define the worst-case risk of a portfolio x as the supremum of $\rho(x,\pi)-\gamma\mu(x,\pi)$ for π that ranges over its uncertainty set. An optimal worst-case portfolio is a composition vector x that minimizes this worst-case risk.

We detail in the paper two numerical schemes that permit us to efficiently evaluate and optimize the worst-case risk in both the variance and the absolute deviation cases. For the variance-based risk measure, the worst-case optimal portfolio can be determined using an interior-point barrier method, in conjunction with an analytic center cutting plane technique. The absolute deviation-based risk measure poses a slight additional complication, due to nonconcavity in π of this function. This issue is here resolved by adding a suitable line search to the algorithm.

The paper is organized as follows. Section 2 sets the stage by providing the basic definitions and introducing the distribution uncertainty model. Section 3 discusses a barrier method for computing the worst-case variance-based risk of a given portfolio, whereas section 4.1 describes the overall cutting plane algorithm for optimizing the worst-case risk over the portfolio composition. Section 5 extends the methodology to the absolute deviation-based risk measure. Some numerical examples are presented in section 6, and conclusions are finally drawn in section 7. To improve readability, some of the technical details have been relegated to appendices.

1.1. Notation. Whenever useful for notational compactness, we use MATLAB-like notation for operations on vectors. If x, y are two vectors of compatible dimensions, relational operators such as $>, \geq$, etc., are to be intended elementwise (e.g., $x > y$ means that all entries of vector $x - y$ are positive). Similarly, powers and operators $+, -, *, /$ work elementwise, and the same holds for standard functions. For example, $\log x/y$ denotes a vector whose ith entry is $\log x_i/y_i$.

2. Preliminaries. Consider a collection of assets or asset classes $a_1, \ldots, a_n$ and let

$$r \doteq \begin{bmatrix} r_1 & \cdots & r_n \end{bmatrix}^\top$$

be a random vector describing the returns of the considered assets over a fixed period of time. Let $r(1), \ldots, r(T)$ be T possible scenarios for the outcomes of the random return vector r, and let π_k be the probability associated to the scenario $r(k)$, with the obvious properties that

$$\pi_k \geq 0, \quad k = 1, \ldots, T,$$
$$\sum_{k=1}^{T} \pi_k = 1.$$

Defining the probability vector

$$\pi \doteq [\pi_1 \ \cdots \ \pi_T]^\top,$$

the two previous conditions are simply rewritten as $\pi \geq 0$, $\mathbf{1}^\top \pi = 1$, where $\mathbf{1}$ denotes a vector of ones of suitable dimensions. Now let

$$x \doteq \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^\top$$

be a vector such that x_i represents the fraction of an investor portfolio that is invested in asset a_i. We shall refer to x as the "portfolio composition," or "portfolio mix."

The portfolio composition can be subject to various kinds of constraints, which we assume to be representable by the condition

$$x \in \mathcal{X}, \quad \text{where } \mathcal{X} \text{ is a given polytope.}$$

For example, a typical form for the set $\mathcal{X}$ is

$$\mathcal{X} = \left\{ x : \sum_{i=1}^{n} x_i = 1,\ x_i \geq 0 \text{ for } i = 1, \ldots, n \right\}, \tag{2.1}$$

which reflects the standard situation where the investor cannot hold a negative amount of an asset (i.e., short-selling is not allowed). However, the results in this paper are not restricted to the specific admissible portfolios set in (2.1) and apply to the general polytopic case.

With the positions above, the investor's total return at the end of the investment period is represented by the random variable

$$w \doteq r^\top x, \tag{2.2}$$

whose expected value is

$$\mu(x, \pi) \doteq \mathrm{E}\left\{ r^\top x \right\} = \sum_{k=1}^{T} \pi_k r^\top(k) x = \left(\sum_{k=1}^{T} \pi_k r^\top(k) \right) x = \hat{r}^\top(\pi) x,$$

where $\hat{r}(\pi) \doteq \mathrm{E}\{r\} = \sum_{k=1}^{T} \pi_k r(k)$.

The portfolio risk is quantified as a measure of variability of w around its expectation. A classical measure of variability (see, e.g., [25]) is given by the *variance*

$$\rho_2(x, \pi) \doteq \mathrm{E}\left\{ \left(r^\top x - \mathrm{E}\left\{ r^\top x \right\} \right)^2 \right\} = x^\top \Sigma(\pi)\, x, \tag{2.3}$$

where

$$\Sigma(\pi) \doteq \mathrm{E}\left\{ (r - \hat{r}(\pi))(r - \hat{r}(\pi))^\top \right\} = \sum_{k=1}^{T} \pi_k (r(k) - \hat{r}(\pi))(r(k) - \hat{r}(\pi))^\top$$

is the covariance matrix of r. In this paper, we also consider an alternative measure of risk, which is based on the expected absolute deviation, and whose justification in the portfolio selection context is discussed, for instance, in [21, 31]:

$$\rho_1(x, \pi) \doteq \mathrm{E}\left\{ \left| r^\top x - \mathrm{E}\left\{ r^\top x \right\} \right| \right\} = \sum_{k=1}^{T} \pi_k |r^\top(k) x - \mu(x, \pi)|.$$

Following a mean-risk approach, we introduce an objective function which represents a tradeoff between risk (variance or expected absolute deviation) and expected return of the portfolio. Specifically, for given $\gamma \geq 0$, we define an objective based on the variance measure

$$\Upsilon_2(x, \pi) \doteq \rho_2(x, \pi) - \gamma \mu(x, \pi) \tag{2.4}$$

and one based on the absolute deviation measure

$$\Upsilon_1(x, \pi) \doteq \rho_1(x, \pi) - \gamma \mu(x, \pi). \tag{2.5}$$

Notice that if the probability distribution π is known and given, then minimizing $\Upsilon_2(x,\pi)$ over $x \in \mathcal{X}$ is a well-known Markowitz problem, whose solution can be obtained by solving numerically a convex quadratic programming problem. Minimizing $\Upsilon_1(x,\pi)$ in this same situation amounts instead to solving a linear programming problem; see, for instance, [21, 31].

The point of this paper is to propose computationally efficient schemes for determining optimal worst-case portfolios, when the probability distribution π is not precisely known. To this end, we introduce in the next section an uncertainty model for π and define the related robust risk functions.

2.1. Distribution ambiguity and robust measures of risk. Assume that a *nominal* return probability distribution η is given, for instance, as a result of estimation from samples. Then the Kullback–Leibler (KL) divergence (see [22]) represents a natural measure of the expected amount of information in a sample from the unknown distribution for discriminating against η [19], and it is a frequently used information-theoretic "distance" measure between probability distributions; see, e.g., [1, 10]. If π, η are two probability vectors in $\mathbb{R}^T$, with $\eta > 0$ describing the nominal probability, the KL distance between π and η is defined as

$$\mathrm{KL}(\pi,\eta) \doteq \sum_{k=1}^{T} \pi_k \log \frac{\pi_k}{\eta_k}.$$

We shall henceforth assume that the "true" probability π is only known to lie within KL distance $d \geq 0$ from η, i.e., $\pi \in \mathcal{K}(\eta,d)$, where

$$\mathcal{K}(\eta,d) \doteq \{\pi \in \Pi : \mathrm{KL}(\pi,\eta) \leq d\},$$

Π being the probability simplex $\Pi = \{\pi : \pi \geq 0,\ \mathbf{1}^\top \pi = 1\}$.

$\mathcal{K}(\eta,d)$ thus represents the *ambiguity set* for the return distribution, and $d \geq 0$ is the uncertainty level (radius of ambiguity). The risk functions (2.4), (2.5), with $\pi \in \mathcal{K}(\eta,d)$, are *ambiguous* risk functions. The nominal distribution η and the ambiguity level may either be assigned by expert advice or estimated from data; see, for instance, [20]. In what follows we shall not investigate further the issue of determination of η and d and shall assume that these quantities are given data.

Remark 1 (domain, range, and convexity of $\mathrm{KL}(\pi,\eta)$). Notice that, since $\pi_k \log \pi_k$ is a convex function over the domain $\pi_k \geq 0$,[1] then $\mathrm{KL}(\pi,\eta)$ is a convex function in π over Π, and hence the uncertainty set $\mathcal{K}(\eta,d)$ is convex.

The function $\mathrm{KL}(\pi,\eta)$, with $\pi \in \Pi$, takes values in the interval $[0,\ \log 1/\eta_{\min}]$, where $\eta_{\min} = \min_{k=1,\ldots,T} \eta_k$. The lower end of the interval is attained for $\pi = \eta$, whereas the higher end is attained for $\pi = e_i$, where i is the index of the smallest element in η, and e_i is the ith vector in the standard basis of $\mathbb{R}^T$.

Figure 2.1 gives a pictorial idea of the shape of the set $\mathcal{K}(\eta,d)$ in a three-dimensional example where η is assumed to be the uniform distribution.

We pursue a worst-case approach in dealing with ambiguity in the risk functions. Notice that it is known (see [30, Theorem 2]) that the risk functions (2.4), (2.5), for a fixed probability π, can be represented in dual form as the result of a maximization

$$\Upsilon(x,\pi) = \max_{\zeta \in A(\pi)} \left[\langle \zeta, r^\top x\rangle_\pi - \Upsilon^*(\zeta)\right],$$

[1] It is assumed by continuity that $0 \log 0 = 0$.

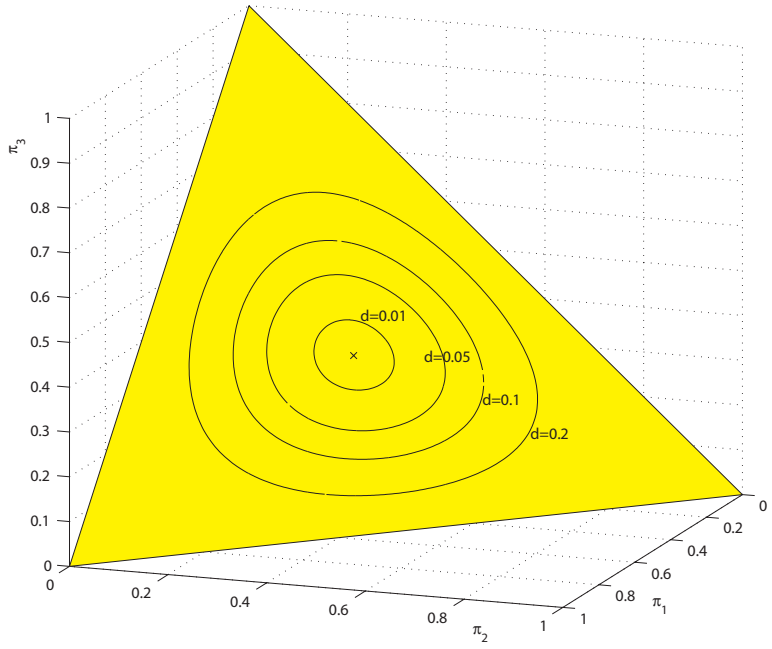


FIG. 2.1. *A visualization of the subsets $\mathcal{K}(\eta, d)$ of the probability simplex in a three-dimensional example with $\eta = [1/3\ 1/3\ 1/3]$ and $d = 0.01,\ 0.05,\ 0.1,\ 0.2$.*

where $A(\pi)$ is a closed convex set of measures, Υ^* is the conjugate of Υ, and $\langle \zeta, w \rangle_\pi = \sum_i \zeta_i w_i \pi_i$. In the the worst-case approach that we follow in this paper, the "robustness" of the nominal risk functions is improved by adding a second level of maximization over a set of admissible probabilities $\mathcal{K}$. That is, we shall consider robust risk functions of the form

$$\Upsilon_{\rm wc}(x) = \max_{\pi \in \mathcal{K}} \Upsilon(x, \pi) = \max_{\pi \in \mathcal{K}} \max_{\zeta \in A(\pi)} \left[\langle \zeta, r^\top x \rangle_\pi - \Upsilon^*(\zeta) \right].$$

Specifically, given the ambiguity model $\mathcal{K}(\eta, d)$ for the return distribution, we define the following worst-case (or robust) measures of risk for a portfolio with composition x:

$$\Upsilon_{\rm wc2}(x) \doteq \max_{\pi \in \mathcal{K}(\eta,d)} \rho_2(x, \pi) - \gamma \mu(x, \pi) \tag{2.6}$$

for the variance-based measure, and

$$\Upsilon_{\rm wc1}(x) \doteq \max_{\pi \in \mathcal{K}(\eta,d)} \rho_1(x, \pi) - \gamma \mu(x, \pi) \tag{2.7}$$

for the absolute deviation-based measure.

The distribution $\pi_{\rm wc}$ that attains the supremum in the above optimization problems is named the *worst-case distribution*, and the corresponding value function $\Upsilon_{\rm wc}(x)$ is the *worst-case risk* (to uncertainty level d) of the portfolio x. In the next section we provide an efficient numerical scheme for solving (2.6). We anticipate that the existence of a polynomial-time algorithm for computing the worst-case variance-based risk is due to the fact that we can construct a self-concordant barrier for the

convex domain $\mathcal{K}(\eta, d)$. Successively, in section 4, we develop a polynomial-time algorithm that permits us to further optimize $\Upsilon_{\text{wc2}}(x)$ with respect to x, and hence to find an optimal portfolio mix that minimizes the worst-case variance-based risk. In section 5, we describe a similar approach for dealing with the absolute deviation-based objective (2.7).

3. Computing the worst-case variance-based risk. Let the portfolio composition x be fixed, let $w(k) = r^\top(k)x$, $k = 1, \ldots, T$, and define

$$\mathbf{w} \doteq \left[\begin{array}{ccc} w(1) & \cdots & w(T) \end{array} \right]^\top.$$

Then from (2.2)–(2.3) we have

$$\rho_2(x, \pi) = \mathrm{E}\left\{(w - \mathrm{E}\{w\})^2\right\} = \mathrm{E}\left\{w^2\right\} - \mathrm{E}^2\{w\} = \pi^\top \mathbf{w}^2 - \pi^\top \Omega \pi,$$

where $\Omega \doteq \mathbf{w}\mathbf{w}^\top$. Since Ω is symmetric positive semidefinite, it follows that $\rho_2(x, \pi)$ is a concave function of the probability vector π. Therefore, the objective function (2.4),

$$\Upsilon_2(x, \pi) = \rho_2(x, \pi) - \gamma\mu(x, \pi) = -\pi^\top \Omega \pi - \pi^\top(\gamma \mathbf{w} - \mathbf{w}^2),$$

is also concave in π; hence problem (2.6) can be written in the equivalent form of a convex minimization problem as follows:

$$\Upsilon_{\text{wc2}} = -\min_{\pi} \pi^\top \Omega \pi + \pi^\top(\gamma \mathbf{w} - \mathbf{w}^2) \tag{3.1}$$

$$\text{subject to } \mathrm{KL}(\pi, \eta) \leq d, \tag{3.2}$$

$$\pi \geq 0, \tag{3.3}$$

$$\mathbf{1}^\top \pi = 1.$$

We next develop an interior-point barrier method for solving problem (3.1).

3.1. A logarithmic barrier method. For a fixed portfolio x, we solve problem (3.1) by solving a sequence of equality constrained problems of the form

$$\min_{\pi} \; f(\pi) \doteq t f_0(\pi) + \phi(\pi) \tag{3.4}$$

$$\text{subject to } \mathbf{1}^\top \pi = 1 \tag{3.5}$$

for increasing values of $t \geq 0$, where

$$f_0(\pi) \doteq \pi^\top \Omega \pi + \pi^\top(\gamma \mathbf{w} - \mathbf{w}^2)$$

is the objective function of (3.1),

$$b(\pi) \doteq \sum_{k=1}^{T} \pi_k \log \pi_k - \sum_{k=1}^{T} \pi_k \log \eta_k - d, \tag{3.6}$$

and

$$\phi(\pi) \doteq -\log\left(-b(\pi)\right) - \sum_{k=1}^{T} \log \pi_k \tag{3.7}$$

is a logarithmic barrier for the inequality constraints (3.2), (3.3). For fixed $t \geq 0$, we denote by $\pi^*(t)$ the corresponding optimal solution of (3.4). The *central path* is the curve $\pi^*(t)$ obtained varying t from 0 to ∞. A standard implementation of a barrier method is hence the following.

ALGORITHM 1 (barrier method [5]).
Given strictly feasible π, set $t = t(0) > 0$, $\varrho > 1$, tolerance $\epsilon > 0$.
repeat

1. Centering step:
   Compute $\pi^*(t)$ by solving (3.4) using the Newton method, starting at π.
2. Update: $\pi = \pi^*(t)$.
3. Stopping criterion: quit if $(T+1)/t < \epsilon$.
4. Increase t: $t = \varrho t$.

Notice that since $\eta > 0$, an initial feasible point for the algorithm is simply given by $\pi = \eta$. For this algorithm, $\pi^*(t)$ tends to the optimal solution of problem (3.1) as $t \to \infty$. The convergence properties of the method are analyzed in terms of the number of *outer iterations* (centering steps) needed to reach a solution with the desired accuracy ϵ and the number of *inner iterations* (i.e., the iterations required by the Newton method to compute each center, up to accuracy $\epsilon_{\rm nw}$). A standard result states that the number of outer iterations (centering steps) is given exactly by (see [5, section 11.3.3])

$$1 + \left\lceil \frac{\log(T+1)/(\epsilon t(0))}{\log \varrho} \right\rceil.$$

The analysis of complexity of each centering step relies on the property of *self-concordance* of the objective function in (3.4), which is discussed next.

3.2. Centering step and self-concordance. In their seminal work [28], Nesterov and Nemirovskii provided a key condition under which the complexity of the Newton method could be analyzed, that is, self-concordance of the objective function. We next show that the objective function in (3.4) is indeed self-concordant and provide a bound on the number of Newton steps required in each centering phase.

We start with some definitions. A function of a scalar variable $\psi(z) : \mathbb{R} \to \mathbb{R}$ is self-concordant if it is convex and

$$|\psi^{(3)}(z)| \leq k \psi^{(2)}(z)^{3/2}$$

for all z in the domain of ψ, where $\psi^{(2)}, \psi^{(3)}$ denote the second and the third derivatives of ψ, respectively, and k is a positive constant. A function $\psi(z)$ of vector variable $z \in \mathbb{R}^n$ is self-concordant if it is self-concordant along any line in its domain, i.e., if the function of scalar variable $\tilde{\psi}(\alpha) \doteq \psi(z + \alpha v)$ is a self-concordant function of $\alpha \in \mathbb{R}$ for all z in the domain of ψ and for all v.

The following proposition on the self-concordance of the barrier function (3.7) holds; see Appendix A for a proof.

PROPOSITION 3.1. *The function $\phi(\pi)$ in* (3.7) *is a self-concordant barrier for the domain*

$$\{(\pi, d) : \ \pi > 0, \varphi(\pi) < d\}.$$

Since the sum of self-concordant functions is self-concordant, and since convex quadratic functions are obviously self-concordant (they have zero third derivative; see also some standard rules of "self-concordant calculus" in [5]), we deduce from

Proposition 3.1 that the objective function in (3.4) is indeed self-concordant. It then follows that using an (equality-constrained) Newton method with backtracking line-search, each center can be computed up to accuracy ϵ_{nw} in at most (see, for instance, [5, section 11])

$$\frac{(T+1)(\varrho - 1 - \log \varrho)}{\ell} + \log_2 \log_2 \frac{1}{\epsilon_{\mathrm{nw}}}$$

Newton steps, where ℓ is a constant that depends on two technical parameters used in the line-search phase of the algorithm. We conclude this section by reporting explicitly the gradient and Hessian of the function $f(\pi)$ in (3.4). We have

$$\begin{aligned} \nabla f(\pi) &= t\nabla f_0(\pi) + \nabla \phi(\pi), \\ \nabla^2 f(\pi) &= t\nabla^2 f_0(\pi) + \nabla^2 \phi(\pi) \end{aligned}$$

with

$$\begin{aligned} \nabla f_0(\pi) &= 2\Omega\pi + (\gamma \mathbf{w} - \mathbf{w}^2), \\ \nabla \phi(\pi) &= \frac{1}{-b(\pi)} \nabla b(\pi) - \pi^{-1}; \quad \nabla b(\pi) = \mathbf{1} + \log \frac{\pi}{\eta} \end{aligned}$$

and

$$\begin{aligned} \nabla^2 f_0(\pi) &= 2\Omega, \\ \nabla^2 \phi(\pi) &= \frac{1}{-b(\pi)} \operatorname{diag}(\pi^{-1}) + \frac{1}{b^2(\pi)} \nabla b(\pi) \nabla^\top b(\pi) + \operatorname{diag}(\pi^{-2}). \end{aligned}$$

4. Minimizing the worst-case variance-based risk. In the previous section, we described a numerically efficient technique for computing the worst-case risk of a *given* portfolio mix x, i.e., for evaluating the function $\Upsilon_{\mathrm{wc2}}(x)$ in (2.6). We now elaborate on this technique and develop an efficient algorithm for determining a portfolio mix that minimizes the worst-case risk. That is, we now aim at solving the portfolio design problem

$$\min_{x \in \mathcal{X}} \ \Upsilon_{\mathrm{wc2}}(x). \tag{4.1}$$

We shall do so by employing an analytic center cutting plane technique, which is described in the next section. Notice preliminarily that function $\Upsilon_2(x, \pi)$ in (2.4),

$$\Upsilon_2(x, \pi) = x^T \Sigma(\pi) x - \gamma \hat{r}^\top(\pi) x,$$

is convex (and quadratic) in x for any given π, whence the function $\Upsilon_{\mathrm{wc2}}(x)$, which is defined as the pointwise maximum of $\Upsilon_2(x, \pi)$ over π, is also convex in x. At any given π, the gradient of $\Upsilon_2(x, \pi)$ with respect to x is

$$\nabla_x \Upsilon_2(x, \pi) = 2\Sigma(\pi) x - \gamma \hat{r}(\pi). \tag{4.2}$$

The gradient defines a supporting hyperplane for the epigraph of $\Upsilon_2(x, \pi)$, i.e.,

$$\Upsilon_2(z, \pi) \geq \Upsilon_2(x, \pi) + [\nabla_x \Upsilon_2(x, \pi)]^\top (z - x) \quad \forall z \in \mathcal{X}. \tag{4.3}$$

Now let x be a given point and let $\pi^*(x)$ be the probability vector that attains the optimal value in problem (2.6) (such an optimal argument is attained, since the feasible set is compact), so that

$$\Upsilon_{\mathrm{wc2}}(x) = \Upsilon_2(x, \pi^*(x)). \tag{4.4}$$

Evaluating (4.3) in $\pi = \pi^*(x)$, we get

$$\Upsilon_2(z, \pi^*(x)) \geq \Upsilon_2(x, \pi^*(x)) + [\nabla_x \Upsilon_2(x, \pi^*(x))]^\top (z - x) \quad \forall z \in \mathcal{X}.$$

Since $\Upsilon_{\rm wc2}(z) \geq \Upsilon_2(z, \pi^*(x))$, continuing the previous inequality on the left and using (4.4), we obtain

$$\Upsilon_{\rm wc2}(z) \geq \Upsilon_{\rm wc2}(x) + [\nabla_x \Upsilon_2(x, \pi^*(x))]^\top (z - x) \quad \forall z \in \mathcal{X}, \tag{4.5}$$

which means that $\nabla_x \Upsilon_2(x, \pi^*(x))$ is a subgradient of $\Upsilon_{\rm wc2}(x)$ at the point x. Notice that each time we solve problem (2.6)—or its equivalent formulation (3.1)—for a given x, we get both the value of $\Upsilon_{\rm wc2}(x)$ and the worst-case probability vector $\pi^*(x)$, and hence (evaluating (4.2) for $\pi = \pi^*(x)$) a subgradient of $\Upsilon_{\rm wc2}(x)$ at x.

4.1. An analytic center cutting plane algorithm for optimizing the portfolio mix. We now briefly describe an analytic center cutting plane (ACCP) method for solving problem (4.1). An overview of ACCP techniques for convex optimization can be found, for instance, in [29].

Let initially $\mathcal{P}_1 = \mathcal{X}$, and compute the analytic center $x^{(1)}$ of $\mathcal{P}_1$. The analytic center of a polytope can be efficiently computed by minimizing a logarithmic barrier via a Newton-type algorithm; see, for instance, [17]. Then solve problem (2.6) to get $\Upsilon_{\rm wc2}(x^{(1)})$, along with the worst-case probability $\pi^*(x^{(1)})$ and a subgradient

$$g_1 \doteq \nabla_x \Upsilon_2(x^{(1)}, \pi^*(x^{(1)}))$$

of $\Upsilon_{\rm wc2}(x)$ at $x = x^{(1)}$. Using inequality (4.5), notice next that for all points in the hyperplane

$$\{z : g_1^\top (z - x^{(1)}) > 0\}$$

we have that $\Upsilon_{\rm wc2}(z) > \Upsilon_{\rm wc2}(x^{(1)})$, hence all such points are worse than the current point $x^{(1)}$ in terms of the objective value that we are trying to minimize. Therefore, the optimal point should lie in the complementary hyperplane

$$\mathcal{H}_1 \doteq \{z : g_1^\top (z - x^{(1)}) \leq 0\}.$$

Hence, we update the current polytope by adding the constraint $\mathcal{H}_1$ to $\mathcal{P}_1$, i.e., we set

$$\mathcal{P}_2 = \mathcal{P}_1 \cap \mathcal{H}_1$$

and iterate the whole process (compute the analytic center $x^{(2)}$ of $\mathcal{P}_2$, etc.).

The convergence of this method relies on the fact that the polytopes $\mathcal{P}_k$ shrink at each iteration, thus eventually localizing the optimal solution x^*. The ACCP method converges to a solution in polynomial time. A precise assessment of the numerical complexity of the ACCP method and some of its variants has been discussed in several papers; see, for instance, [16, 17].

In a practical implementation of the method, we may terminate the iterations if either $\|x^{(k)} - x^{(k-1)}\|$ goes below a given threshold $\epsilon_{\rm ac}$ or the Chebyshev radius of $\mathcal{P}_k$ becomes sufficiently small.[2]

[2]The Chebyshev radius of a polytope is defined as the radius of the largest Euclidean hypersphere contained in the polytope. Computing the Chebyshev radius amounts to solving a linear programming problem; therefore checking the exit condition based on the Chebyshev radius requires some additional numerical effort.

A schematic implementation of an algorithm that permits us to solve (up to a numerical tolerance) the robust portfolio design problem is given next.

ALGORITHM 2 (ACCP).
Given the exit tolerance $\epsilon_{\text{ac}} > 0$, and the initial polytope $\mathcal{X}$, set $\mathcal{P} = \mathcal{X}$.
repeat

1. Centering step:
   Compute the analytic center x of $\mathcal{P}$.
2. Solve subproblem (2.6):
   Compute $\Upsilon_{\text{wc2}}(x)$, $\pi^*(x)$, and a subgradient g of Υ_{wc2} at x.
3. Stopping criterion:
   If Chebyshev-radius$(\mathcal{P}) < \epsilon_{\text{ac}}$, then quit.
4. Update the polytope:
   Set $\mathcal{P} = \mathcal{P} \cap \{z : g^\top (z - x) \leq 0\}$.

5. The worst-case absolute deviation-based risk. The robust design approach outlined in the previous section for the variance-based measure can be extended to the absolute deviation-based measure (2.7). In this section, we mainly discuss how to evaluate the worst-case absolute deviation risk of a given portfolio x,

$$\Upsilon_{\text{wc1}}(x) \doteq \max_{\pi \in \mathcal{K}(\eta,d)} \Upsilon_1(x,\pi), \tag{5.1}$$

and then hint at how to minimize $\Upsilon_{\text{wc1}}(x)$ over $x \in \mathcal{X}$ in section 5.1.1. This latter process is completely analogous to the one described for the variance-based risk function.

The main technical difference with respect to the case considered previously is that, contrary to the variance function $\rho_2(x,\pi)$, the absolute deviation function $\rho_1(x,\pi)$ is *not* concave in π; see, e.g., Figure 5.1, and Appendix B for a proof. Therefore, the inner problem (5.1) cannot be solved by directly using a logarithmic barrier method such as the one described in section 3.1. However, we show in the next section that (5.1) can still be solved efficiently to any given accuracy by using the barrier method in conjunction with a one-dimensional search.

5.1. Evaluating the worst-case absolute deviation risk. We assume the portfolio composition x to be fixed, and use the notation introduced in section 3. The absolute measure and objective function in (5.1) are then

$$\rho_1(\pi) \doteq \rho_1(x,\pi) = \sum_{k=1}^{T} \pi_k |w(k) - \mathbf{w}^\top \pi|, \tag{5.2}$$

$$\Upsilon_1(\pi) \doteq \Upsilon_1(x,\pi) = \sum_{k=1}^{T} \pi_k |w(k) - \mathbf{w}^\top \pi|, -\gamma \mathbf{w}^\top \pi. \tag{5.3}$$

As we mentioned before, the function $\rho_1(x,\pi)$, and hence $\Upsilon_1(x,\pi)$, is not concave (nor convex) over $\pi \in \Pi$. However, $\Upsilon_1(x,\pi)$ is concave (and actually linear) in π if we fix the value of the expected value: $\mathbf{w}^\top \pi \overset{!}{=} \mu$.

The solution idea is therefore the following one. First, determine the extreme feasible values for the portfolio mean return:

$$\mu_{\min} = \min_{\pi \in \mathcal{K}(\eta,d)} \mathbf{w}^\top \pi; \quad \mu_{\max} = \max_{\pi \in \mathcal{K}(\eta,d)} \mathbf{w}^\top \pi \tag{5.4}$$

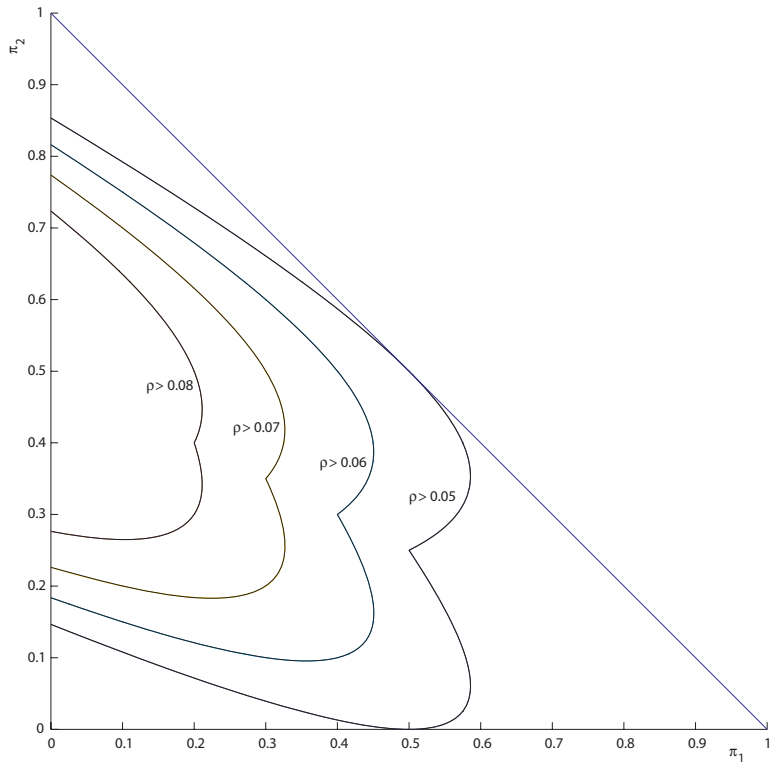


FIG. 5.1. *Projection of the superlevel sets of* (5.2) *on the* π_1, π_2 *plane in an example with* $T = 3$ *and* $\mathbf{w} = [0.1\ 0.2\ 0.3]^\top$.

(we show in Appendix C that these two values can be computed very quickly via a scalar bisection algorithm). Then, for $\mu \in [\mu_{\min},\ \mu_{\max}]$, define

$$\varphi(\mu) \doteq \max_{\pi \in \mathcal{K}(\eta,d)} \sum_{k=1}^{T} \pi_k |w(k) - \mu| - \gamma\mu \quad \text{subject to } \mathbf{w}^\top \pi = \mu. \tag{5.5}$$

Clearly, we have that

$$\Upsilon_{\mathrm{wc1}}(x) = \max_{\pi \in \mathcal{K}(\eta,d)} \Upsilon_1(x,\pi) = \max_{\mu \in [\mu_{\min},\ \mu_{\max}]} \varphi(\mu). \tag{5.6}$$

In practice, we divide the interval $[\mu_{\min},\ \mu_{\max}]$ into N grid points $\mu_1, \ldots, \mu_N$, where N is chosen in accordance to the desired solution accuracy. For $i = 1, \ldots, N$, computing $\varphi(\mu_i)$ is a convex optimization program that can be solved efficiently using, for instance, a barrier method such as the one described in section 3.1. An approximate solution to (5.1) is hence given by

$$\Upsilon_{\mathrm{wc1}}(x) \simeq \max_{i=1,\ldots,N} \varphi(\mu_i).$$

Notice that the main difficulty in (5.5) is due to the presence of the KL constraint. It is instructive to detail the solution of (5.5) in the particular situation when this

constraint is not present, since a closed-form solution is obtained in this case. The solution in (5.7) is computed basically at no cost, and it is optimal for problem (5.5) if the constraint $\mathrm{KL}(\pi, \eta) \le d$ happens to be inactive.

PROPOSITION 5.1. *Assume without loss of generality that the values* $w(1), \ldots, w(T)$ *are arranged in increasing order, and let* π *be such that*

$$\pi_1 = \frac{w(T) - \mu}{w(T) - w(1)}, \quad \pi_2 = 0, \ \ldots, \ \pi_{T-1} = 0, \quad \pi_T = \frac{\mu - w(1)}{w(T) - w(1)}. \tag{5.7}$$

If $\mathrm{KL}(\pi, \eta) \le d$, *then* π *is an optimal solution for problem* (5.5), *with corresponding optimal value function*

$$\varphi(\mu) = 2\frac{-\mu^2 + \mu(w(1) + w(T)) - w(1)w(T)}{w(T) - w(1)} - \gamma\mu. \tag{5.8}$$

A proof of this proposition is given in Appendix D.

5.1.1. Optimizing over the portfolio composition. The procedure described in the previous section can further be wrapped by a cutting plane scheme, similarly to the one described in section 4, in order to optimize $\Upsilon_{\mathrm{wc1}}(x)$ over the portfolio mix x.

Notice that the absolute deviation measure (2.5) is convex in x for any given π. A subgradient of $\Upsilon_1(x, \pi)$ at point x is given by

$$g_x(x, \pi) = \sum_{k=1}^{T} \pi_k (r(k) - \hat{r}(\pi)) s_k - \gamma \hat{r}(\pi),$$

where

$$s_k \doteq \begin{cases} 1 & \text{if } (r(k) - \hat{r})x \ge 0, \\ -1 & \text{otherwise.} \end{cases}$$

Now let $\pi^*(x)$ denote the probability of attaining the optimum in problem (5.6): following steps similar to (4.2)–(4.5) we have that $g_x(x, \pi^*(x))$ is a subgradient of $\Upsilon_{\mathrm{wc1}}(x)$ at x. This subgradient can be used in the cutting plane scheme of section 4.1, thus providing an overall polynomial-time method to solve the worst-case design problem $\min_{x \in \mathcal{X}} \Upsilon_{\mathrm{wc1}}(x)$.

6. Numerical examples. We considered a financial allocation problem over five asset classes, where each class is represented by a sector index. We used the following indices to represent the classes: (1) Russell 1000 Large Cap Growth Index (RKGR), (2) Russell 1000 Large Cap Value Index (RKVA), (3) Russell 2000 Small Cap Growth Index (R2KGR), (4) Russell 2000 Small Cap Value Index (R2KVA), (5) Merrill Lynch Intermediate Bond Index (MACTX), with historical data of daily logarithmic returns collected over the period from July 14, 2004 to December 30, 2005 ($T = 371$ scenarios). We assumed that the return on the next day after the observed period can take on any of the historical values, with equal probability. This amounts to choosing a uniform nominal distribution η on the scenarios, which also conforms to the approach undertaken in [21, 31].

Return-risk analysis. We first analyze a fixed portfolio x_{fix} which allocates 30% of the wealth in bonds, and the rest equally distributed among the remaining assets. The nominal expected return for this portfolio is $\mu(x_{\mathrm{fix}}, \eta) = 3.8782 \times 10^{-4}$, the nominal

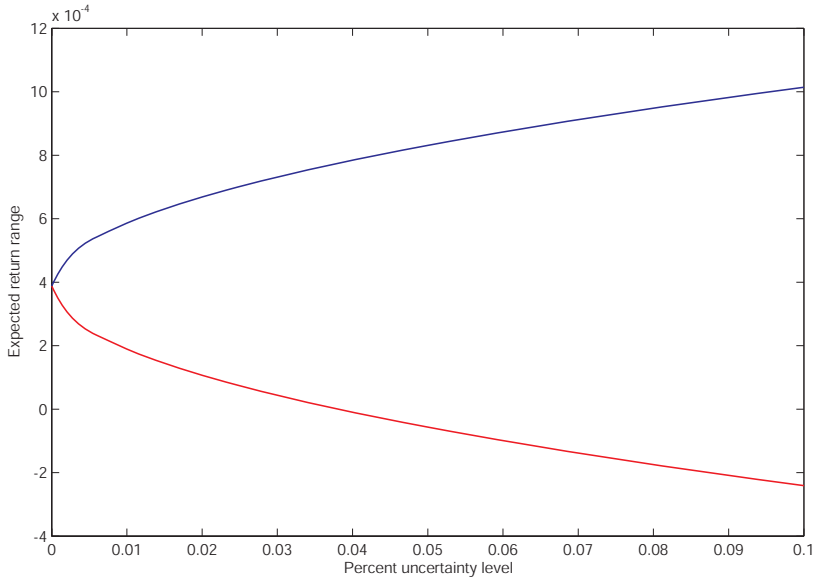


FIG. 6.1. *Upper and lower limits for the expected return of portfolio $x_{\rm fix}$ as a function of the percent uncertainty level (%u.l.).*

variance is $\rho_2(x_{\rm fix}, \eta) = 3.3307 \times 10^{-5}$, and the nominal absolute deviation risk is $\rho_1(x_{\rm fix}, \eta) = 0.0046$.

We can perform various worst-case analyses on this portfolio. First, we computed the range of variation of the expected return, using the technique discussed in Appendix C. The results are shown in Figure 6.1 (the uncertainty level in the plots is expressed in percent units of the maximum allowable value of d, i.e., $\%u.l. = 100\frac{d}{\log 1/\eta_{\min}}$).

Next, we evaluated the worst-case variance-based risk (2.6) of portfolio $x_{\rm fix}$, with $\gamma = 0.1$ and for increasing values of the percent uncertainty level; see Figure 6.2. An analogous plot, obtained from the absolute deviation-based risk measure (2.7) is instead shown in Figure 6.3. Notice from these plots that relatively low uncertainty levels may induce significant variations in the risk measure, with respect to the nominal (no uncertainty) situation.

Return-risk optimization. We next tested the ACCP algorithm described in section 4.1 for optimizing the worst-case variance-based risk. We computed worst-case optimal portfolios at different levels of uncertainty, which resulted in the plot shown in Figure 6.4. The composition of the worst-case optimal portfolios is shown in Figure 6.5.

Numerical performance. In the previous numerical tests, based on nonoptimized codes run under MATLAB 7.2 on an AMD Opteron 280 workstation, we experienced times of less than one minute to compute a worst-case optimal portfolio to an $\epsilon_{\rm ac} = 10^{-5}$ accuracy.

As a further example, we considered the 30 assets composing the Dow Jones Industrial Average Index (DJI) and collected $T = 138$ historical daily return scenarios from March 24, 2006 to October 9, 2006. We ran the variance-based ACCP algorithm on these data, setting $\gamma = 0.2$, $\%u.l. = 0.1$, and exit accuracy $\epsilon_{\rm ac} = 10^{-5}$. Algorithm 2 executed 82 iterations before returning the optimal portfolio. The total execution time was 23.9 seconds. Figure 6.6 shows the reduction in the Chebyshev radius of the

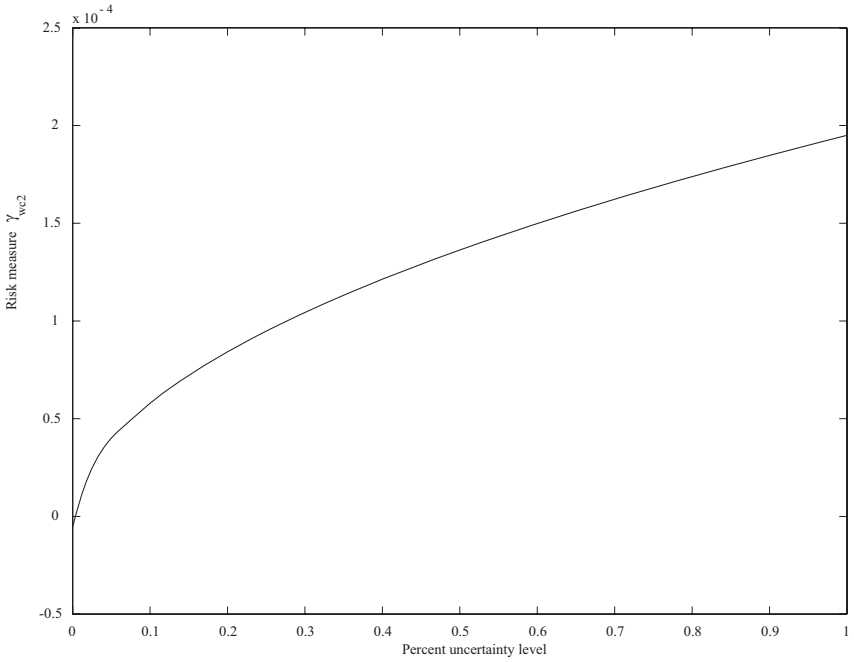


FIG. 6.2. *Worst-case variance-based risk measure of portfolio* x_{fix} *as a function of* %*u.l.*

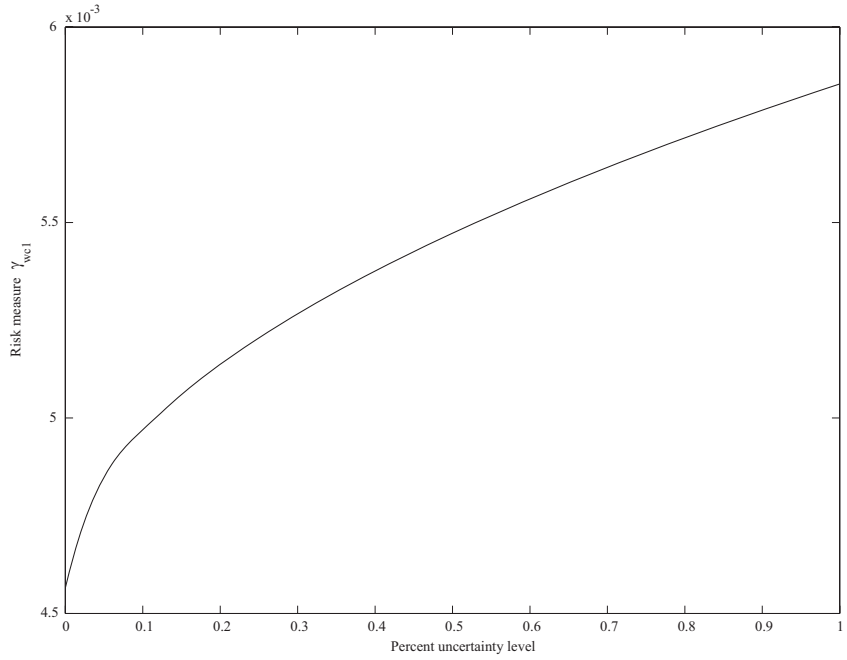


FIG. 6.3. *Worst-case absolute deviation-based risk measure of portfolio* x_{fix} *as a function of* %*u.l.*

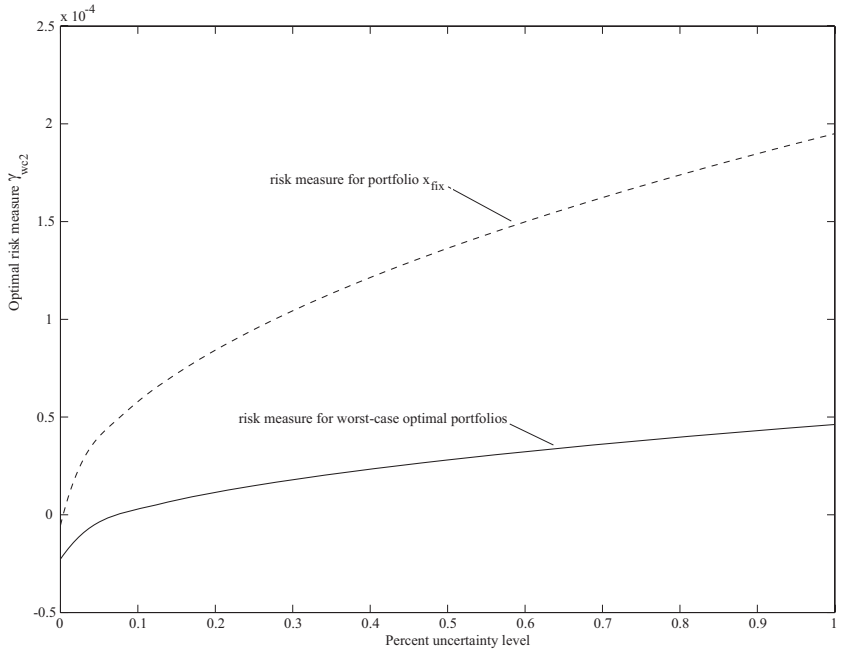


FIG. 6.4. *Dashed line: risk measure of portfolio* $x_{\rm fix}$ *(same as in Figure* 6.2*). Solid line: risk measure for optimal portfolios.*

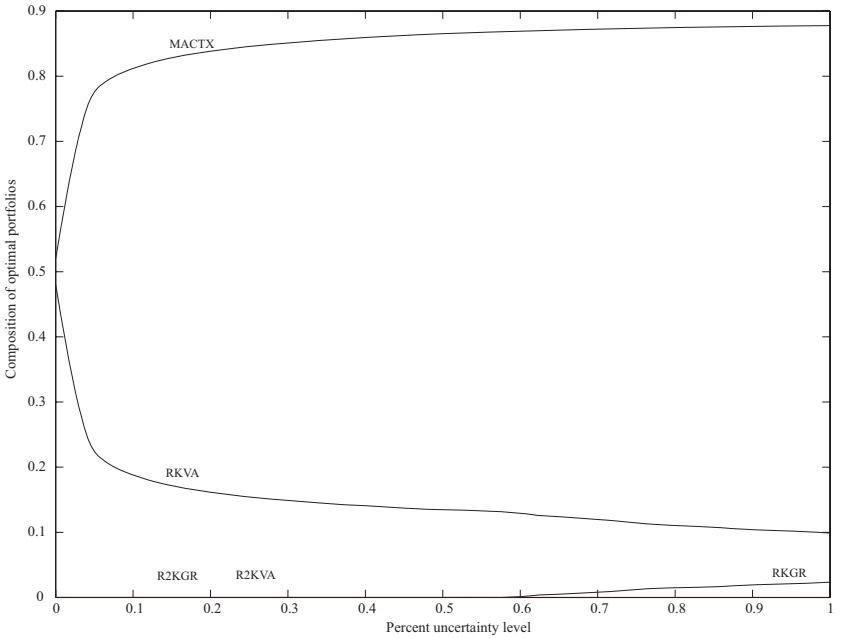


FIG. 6.5. *Composition of worst-case optimal portfolios.*

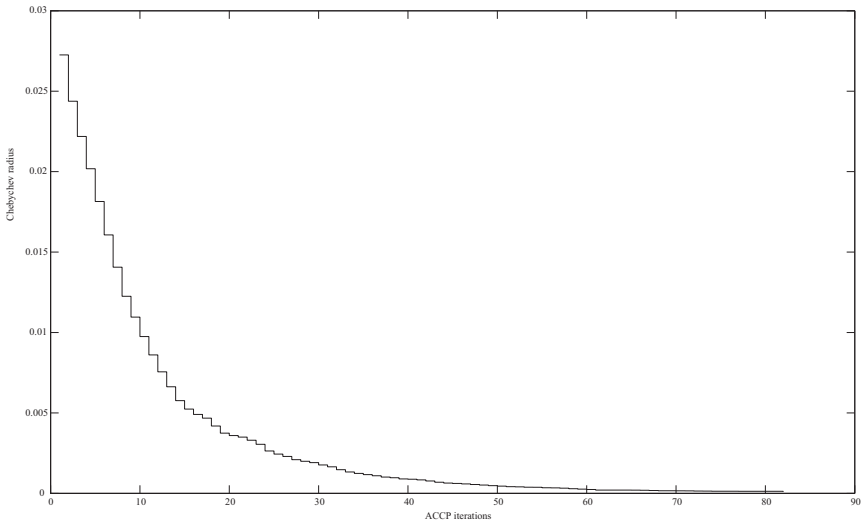


FIG. 6.6. *Chebyshev radius of the localization polytope versus iteration count for computing an optimal robust portfolio in the DJI example.*

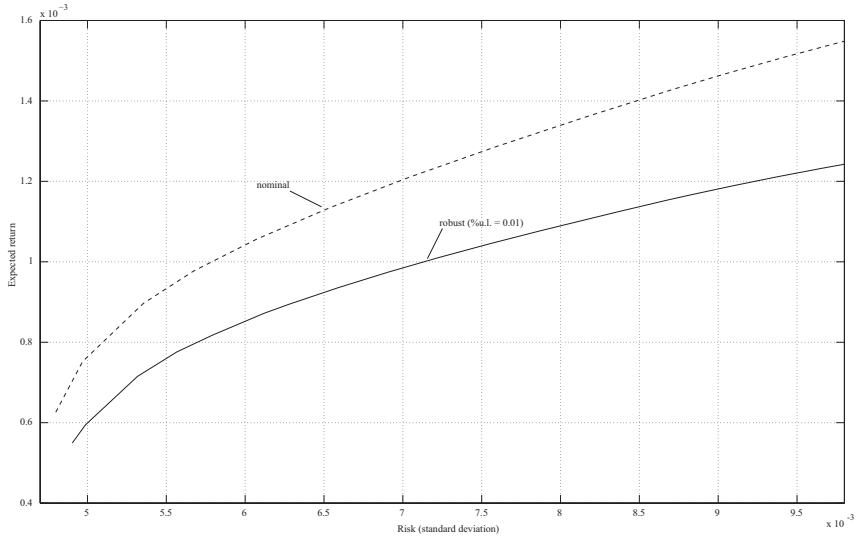


FIG. 6.7. *Nominal efficient frontier for DJI data (dashed), and robust version (solid) for ambiguity level* $\%u.l. = 0.01$.

localization polytope versus the iteration count.

Finally, Figure 6.7 shows a discrete approximation of a portion of the nominal and robust efficient frontiers for the DJI data. The frontiers have been computed at 20 discretized values of $\gamma \in [0.005, 0.3]$. The dashed curve plots the expected return and standard deviation of efficient portfolios under the nominal distribution η (uniform). The solid curve plots the expected return and standard deviation of efficient portfolios under the worst-case distributions (each value of γ results in a different worst-case distribution) for percent ambiguity level $\%u.l. = 0.01$. The worst-case curve was obtained in about 8.5 minutes.

7. Conclusions. Following the recent stream in the literature dealing with statistical model uncertainty (ambiguity) in asset allocation problems, this work explores the case where the ambiguity level in a discrete return distribution is measured according to the Kullback–Leibler divergence. A methodology is proposed for assessing and optimizing the worst-case risk of a portfolio under this type of uncertainty. Two standard risk measures (expected return composed with variance or absolute deviation) are examined, and polynomial-time algorithms are developed for solving efficiently the ensuing problems.

The proposed algorithms are based on interior-point barrier methods for convex optimization, in conjunction with a cutting plane technique. Although it is known in general that cutting plane methods have quite a high iterations-per-digit ratio, they do provide polynomial-time guaranteed convergence to the global optimum to any given practical accuracy. Moreover, in the specific application area discussed in this paper, they permit one to decouple a portfolio analysis phase (step 2 in Algorithm 2) from a mix optimization one (step 1). The numerical experiments show that worst-case optimal portfolios can be computed in reasonable time on a modern computer, and suggest that the proposed methods may be potentially useful in practice for controlling analytically the effects of model uncertainty on financial risk.

Appendix A. Proof of Proposition 3.1. Consider the function

$$\vartheta(\pi) = \sum_{k=1}^{T} \pi_k \log \pi_k - \sum_{k=1}^{T} \pi_k \log \eta_k,$$

which is convex and three times differentiable over the domain $\pi > 0$, and let $D^2\vartheta(\pi)[h,h]$, $D^3\vartheta(\pi)[h,h,h]$ denote, respectively, the second and third differentials of $\vartheta(\pi)$, taken at π along the direction $h \in \mathbb{R}^T$. That is,

$$D^2\vartheta(\pi)[h,h] = \frac{\mathrm{d}^2}{\mathrm{d}t^2}\vartheta(\pi+th)\bigg|_{t=0} = \sum_{k=1}^{T} \frac{h_k^2}{\pi_k},$$

$$D^3\vartheta(\pi)[h,h,h] = \frac{\mathrm{d}^3}{\mathrm{d}t^3}\vartheta(\pi+th)\bigg|_{t=0} = \sum_{k=1}^{T} \frac{-h_k^3}{\pi_k^2}.$$

We have that

$$\begin{aligned}
\left|D^3\vartheta(\pi)[h,h,h]\right| &\le \sum_{k=1}^{T} \frac{|h_k|^3}{\pi_k^2} = \sum_{k=1}^{T} \frac{h_k^2}{\pi_k} \cdot \frac{|h_k|}{\pi_k} \\
&\le \sqrt{\sum_{k=1}^{T} \left(\frac{h_k^2}{\pi_k}\right)^2} \cdot \sqrt{\sum_{k=1}^{T} \left(\frac{|h_k|}{\pi_k}\right)^2} \\
&\le \sum_{k=1}^{T} \frac{h_k^2}{\pi_k} \cdot \sqrt{\sum_{k=1}^{T} \left(\frac{|h_k|}{\pi_k}\right)^2} \\
&= D^2\vartheta(\pi)[h,h] \cdot \sqrt{\sum_{k=1}^{T} \frac{h_k^2}{\pi_k^2}},
\end{aligned}$$

where the first inequality in the chain is the triangle inequality, the second is Hölder's inequality, and the third follows from the inequality $\|x\|_2 \le \|x\|_1$ between the ℓ_2 and

ℓ_1 norms. Summarizing, the relation

$$\left|D^3\vartheta(\pi)[h,h,h]\right| \le D^2\vartheta(\pi)[h,h]\cdot\sqrt{\sum_{k=1}^{T}\frac{h_k^2}{\pi_k^2}}$$

holds for all $\pi > 0$ and $h \in \mathbb{R}^T$. We now apply a known result on logarithmic barriers: Due to the previous inequality, function $\vartheta(\pi)$ satisfies the hypotheses of Lemma 2 of [9], from which it follows that the function

$$-\log(d-\vartheta(\pi)) - \sum_{k=1}^{T}\log\pi_k$$

(which coincides with function $\phi(\pi)$ defined in (3.7)) is a self-concordant barrier for the domain $\{(\pi,d):\ \pi>0, \vartheta(\pi)<d\}$, thus concluding the proof. □

Appendix B. Nonconcavity of the absolute deviation function. Assume without loss of generality that the data $w(k)$ are arranged in increasing order. The function

$$\rho_1(\pi) = \sum_k \pi_k |w(k) - \mathbf{w}^\top\pi|$$

is not only nonconcave over the simplex, but it is nonconcave also on the restricted domains

$$R_i \doteq \{\pi\in\Pi : \mathbf{w}^\top\pi \in (w(i),\, w(i+1))\}.$$

For $\pi\in R_i$, we have that $w(k)-\mathbf{w}^\top\pi>0$ for $k\in K_+ \doteq \{i+1,\dots,T\}$ and $w(k)-\mathbf{w}^\top\pi<0$ for $k\in K_- \doteq \{1,\dots,i\}$. Hence, for $\pi\in R_i$, we may write

$$\begin{aligned}\rho_1(\pi) &= \sum_{k\in K_+}\pi_k(w(k)-\mathbf{w}^\top\pi) - \sum_{k\in K_-}\pi_k(w(k)-\mathbf{w}^\top\pi)\\ &= \sum_{k\in K_+}\pi_k w(k) - \sum_{k\in K_-}\pi_k w(k) - \mathbf{w}^\top\pi\left(\sum_{k\in K_+}\pi_k - \sum_{k\in K_-}\pi_k\right).\end{aligned}$$

From the latter expression, since $\sum_{k\in K_-}\pi_k = 1-\sum_{k\in K_+}\pi_k$ and $\sum_{k\in K_-}\pi_k w(k) = \mathbf{w}^\top\pi - \sum_{k\in K_+}\pi_k w(k)$, we further have

$$\rho_1(\pi) = 2\sum_{k\in K_+}\pi_k(w(k)-\mathbf{w}^\top\pi).$$

Now let $\pi^a,\pi^b\in R_i$ and consider

$$\begin{aligned}\rho_1\left(\frac{1}{2}(\pi^a+\pi^b)\right) &= 2\sum_{k\in K_+}\frac{\pi_k^a+\pi_k^b}{2}\left(w(k)-\mathbf{w}^\top\frac{\pi^a+\pi^b}{2}\right)\\ &= \frac{1}{2}\sum_{k\in K_+}\pi_k^a(w(k)-\mathbf{w}^\top\pi^a) + \frac{1}{2}\sum_{k\in K_+}\pi_k^b(w(k)-\mathbf{w}^\top\pi^b)\\ &\quad+\frac{1}{2}\sum_{k\in K_+}\pi_k^a(w(k)-\mathbf{w}^\top\pi^b) + \frac{1}{2}\sum_{k\in K_+}\pi_k^b(w(k)-\mathbf{w}^\top\pi^a)\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2}\sum_{k\in K_+} \pi_k^a (w(k) - \mathbf{w}^\top \pi^a) + \frac{1}{2}\sum_{k\in K_+} \pi_k^b (w(k) - \mathbf{w}^\top \pi^b) \\
&+ \frac{1}{2}\sum_{k\in K_+} \pi_k^a (w(k) - \mathbf{w}^\top \pi^a) + \frac{1}{2}(\mathbf{w}^\top \pi^a - \mathbf{w}^\top \pi^b) \sum_{k\in K_+} \pi_k^a \\
&+ \frac{1}{2}\sum_{k\in K_+} \pi_k^b (w(k) - \mathbf{w}^\top \pi^b) - \frac{1}{2}(\mathbf{w}^\top \pi^a - \mathbf{w}^\top \pi^b) \sum_{k\in K_+} \pi_k^b \\
&= \frac{1}{2}\rho_1(\pi^a) + \frac{1}{2}\rho_1(\pi^b) + \frac{1}{2}(\mathbf{w}^\top \pi^a - \mathbf{w}^\top \pi^b) \sum_{k\in K_+} (\pi_k^a - \pi_k^b).
\end{aligned}
$$

The last term in the sum is sign-indefinite (in particular, it does not hold in general that $\mathbf{w}^\top \pi^a - \mathbf{w}^\top \pi^b > 0$ implies $\sum_{k\in K_+} \pi_k^a - \sum_{k\in K_+} \pi_k^b > 0$); hence ρ_1 is not concave over the domain R_i. To see this latter point, take, for instance, π^a all zero, except for $\pi_i^a = \pi_{i+1}^a = 1/2$, and π^b all zero, except for $\pi_{i-1}^b = 1/2 - \epsilon$, $\pi_{i+1}^b = 1/2 + \epsilon$, with

$$
\max\left(0,\, \epsilon_{\max} - \frac{1}{2}\frac{w(i+1) - w(i)}{w(i+1) - w(i-1)}\right) < \epsilon < \epsilon_{\max}; \quad \epsilon_{\max} \doteq \frac{1}{2}\frac{w(i) - w(i-1)}{w(i+1) - w(i-1)}.
$$

Then one can check by direct inspection that $\pi^a, \pi^b \in R_i$, $\mathbf{w}^\top \pi^a - \mathbf{w}^\top \pi^b > 0$, but $\sum_{k\in K_+} \pi_k^a - \sum_{k\in K_+} \pi_k^b < 0$.

Appendix C. Computing the range for the expected return. We discuss here an efficient technique for determining the extreme values (5.4) of the expected return of a given portfolio. Consider the problem

$$
\begin{aligned}
\mu_{\min} \doteq \min_{\pi} \quad & \mathbf{w}^\top \pi \\
\text{subject to } & \pi \in \mathcal{K}(\eta, d).
\end{aligned}
\tag{C.1}
$$

The Lagrangian for this problem is

$$
\mathcal{L}(\pi, \lambda_{(\pi)}, \lambda_{(\mathrm{kl})}, \nu_{(1)}) = \mathbf{w}^\top \pi - {\lambda_{(\pi)}}^\top \pi + \lambda_{(\mathrm{kl})}(\mathrm{KL}(\pi, \eta) - d) + \nu_{(1)}(\mathbf{1}^\top \pi - 1),
$$

where $\lambda_{(\pi)}, \lambda_{(\mathrm{kl})}, \nu_{(1)}$ are Lagrange multipliers (dual variables). We assume henceforth that $\lambda_{(\mathrm{kl})}$ is strictly positive.[3] The Lagrange dual function is

$$
\begin{aligned}
&g(\lambda_{(\pi)}, \lambda_{(\mathrm{kl})}, \nu_{(1)}) = \inf_{\pi} \mathcal{L}(\pi, \lambda_{(\pi)}, \lambda_{(\mathrm{kl})}, \nu_{(1)}) \\
&= -\lambda_{(\mathrm{kl})} d - \nu_{(1)} + \inf_{\pi} \left(\mathbf{q}^\top \pi + \lambda_{(\mathrm{kl})}\, \mathrm{KL}(\pi, \eta)\right),
\end{aligned}
$$

where

$$
\mathbf{q} \doteq \mathbf{w} - \lambda_{(\pi)} + \nu_{(1)}\mathbf{1}.
$$

Observe now that

$$
\nabla_\pi \left(\mathbf{q}^\top \pi + \lambda_{(\mathrm{kl})}\, \mathrm{KL}(\pi, \eta)\right) = \mathbf{q} + \lambda_{(\mathrm{kl})}(\mathbf{1} + \log \pi/\eta) = 0
$$

for

$$
\pi_k = \eta_k \mathrm{e}^{-\mathbf{q}_k/\lambda_{(\mathrm{kl})} - 1},
$$

[3] When the KL constraint is inactive at optimum, the optimal value of the dual variable $\lambda_{(\mathrm{kl})}$ is zero, due to the complementary slackness condition. However, in this case the solution to problem (C.1) is trivially given by $\min_k w(k)$.

which yields

$$\inf_{\pi}\left(\mathbf{q}^{\top}\pi+\lambda_{(\mathrm{kl})}\,\mathrm{KL}(\pi,\eta)\right)=-\lambda_{(\mathrm{kl})}\sum_{k}\eta_k\mathrm{e}^{-\mathbf{q}_k/\lambda_{(\mathrm{kl})}-1},$$

and hence the dual function

$$g(\lambda_{(\pi)},\lambda_{(\mathrm{kl})},\nu_{(1)})=-\lambda_{(\mathrm{kl})}d-\nu_{(1)}-\lambda_{(\mathrm{kl})}\sum_{k}\eta_k\mathrm{e}^{-\mathbf{q}_k/\lambda_{(\mathrm{kl})}-1}.$$

The dual to problem (C.1) is therefore

$$\begin{aligned}\mu_{\min}\doteq\ &\max -\lambda_{(\mathrm{kl})}d-\nu_{(1)}-\lambda_{(\mathrm{kl})}\sum_{k}\eta_k\mathrm{e}^{-\mathbf{q}_k/\lambda_{(\mathrm{kl})}-1}\\ \text{subject to }&\lambda_{(\mathrm{kl})}>0,\\ &\lambda_{(\pi)}\geq 0\end{aligned}$$

or, equivalently,

$$\begin{aligned}\mu_{\min}\doteq\ &-\min\ \lambda_{(\mathrm{kl})}d+\nu_{(1)}+\lambda_{(\mathrm{kl})}\sum_{k}\eta_k\mathrm{e}^{-\mathbf{q}_k/\lambda_{(\mathrm{kl})}-1}\\ \text{subject to }&\lambda_{(\mathrm{kl})}>0,\\ &\lambda_{(\pi)}\geq 0.\end{aligned}\tag{C.2}$$

Notice that, for fixed $\lambda_{(\mathrm{kl})},\lambda_{(\pi)}$, at the optimum the derivative of (C.2) with respect to $\nu_{(1)}$ must be zero, i.e.,

$$1+\lambda_{(\mathrm{kl})}\sum_{k}\eta_k\mathrm{e}^{-\mathbf{q}_k/\lambda_{(\mathrm{kl})}-1}\frac{-1}{\lambda_{(\mathrm{kl})}}\frac{\partial\mathbf{q}_k}{\partial\nu_{(1)}}=1-\sum_{k}\eta_k\mathrm{e}^{-\mathbf{q}_k/\lambda_{(\mathrm{kl})}-1}=0,$$

from which we obtain

$$\begin{gathered}\mathrm{e}^{-\nu_{(1)}/\lambda_{(\mathrm{kl})}}\sum_{k}\eta_k\mathrm{e}^{(\lambda_{(\pi),\mathrm{k}}-w(k))/\lambda_{(\mathrm{kl})}-1}=1\\ \Downarrow\\ \nu_{(1)}=\lambda_{(\mathrm{kl})}\left(\log\sum_{k}\eta_k\mathrm{e}^{(\lambda_{(\pi),\mathrm{k}}-w(k))/\lambda_{(\mathrm{kl})}}-1\right).\end{gathered}$$

Substituting this latter expression into (C.2), we write the dual in the following reduced form:

$$\begin{aligned}-\mu_{\min}\doteq&\min\lambda_{(\mathrm{kl})}d+\lambda_{(\mathrm{kl})}\log\sum_{k}\eta_k\mathrm{e}^{(\lambda_{(\pi),\mathrm{k}}-w(k))/\lambda_{(\mathrm{kl})}}\\ \text{subject to }&\lambda_{(\mathrm{kl})}>0,\\ &\lambda_{(\pi)}\geq 0.\end{aligned}$$

Observe further that, for any given $\lambda_{(\mathrm{kl})}>0$, the optimal choice for $\lambda_{(\pi)}$ is zero, which finally yields the dual in the form of a univariate convex problem:

$$\begin{aligned}-\mu_{\min}\doteq\min f_l(\lambda_{(\mathrm{kl})})\doteq&\lambda_{(\mathrm{kl})}d+\lambda_{(\mathrm{kl})}\log\sum_{k}\eta_k\mathrm{e}^{-\frac{w(k)}{\lambda_{(\mathrm{kl})}}}\\ \text{subject to }&\lambda_{(\mathrm{kl})}>0.\end{aligned}$$

An identical reasoning can be applied for computing the upper limit $\mu_{\max}$ of the mean range. In this case, we obtain

$$\mu_{\max} \doteq \min f_u(\lambda_{(\mathrm{kl})}) \doteq \lambda_{(\mathrm{kl})} d + \lambda_{(\mathrm{kl})} \log \sum_k \eta_k \mathrm{e}^{\frac{w(k)}{\lambda_{(\mathrm{kl})}}} \quad \text{subject to } \lambda_{(\mathrm{kl})} > 0. \tag{C.3}$$

Both problems can be readily solved via a bisection scheme (such as the one sketched in Algorithm 3 below), given the gradient of the objective function

$$g_l(\lambda) \doteq \frac{\partial f_l}{\partial \lambda} = \frac{1}{\lambda}\left(f_l(\lambda) + \frac{\sum_k w(k)\eta_k \mathrm{e}^{-w(k)/\lambda}}{\sum_k \eta_k \mathrm{e}^{-w(k)/\lambda}} \right)$$

for the lower mean limit, and

$$g_u(\lambda) \doteq \frac{\partial f_u}{\partial \lambda} = \frac{1}{\lambda}\left(f_u(\lambda) - \frac{\sum_k w(k)\eta_k \mathrm{e}^{-w(k)/\lambda}}{\sum_k \eta_k \mathrm{e}^{w(k)/\lambda}} \right)$$

for the upper limit.

ALGORITHM 3 (bisection).
Given initial $\lambda_l = 0$, $\lambda_r = 1$ and tolerance $\epsilon > 0$

1. while $g(\lambda_r) < 0$, let $\lambda_r = 2\lambda_r$, end while;
2. while $\lambda_r - \lambda_l > \epsilon$
   2.a let $\lambda = \frac{1}{2}(\lambda_r + \lambda_l)$
   2.b if $g(\lambda) > 0$, let $\lambda_r = \lambda$; else let $\lambda_l = \lambda$, end if
3. end while
4. return λ.

A perhaps interesting observation is that the function appearing in (C.3)

$$\lambda_{(\mathrm{kl})} \log \sum_k \eta_k \mathrm{e}^{\frac{w(k)}{\lambda_{(\mathrm{kl})}}}$$

tends to $\max_k w(k)$ as $\lambda_{(\mathrm{kl})} \to 0$. Indeed, for $\lambda_{(\mathrm{kl})} > 0$, this function is a "uniform smooth approximation" of the max function, using the terminology introduced in [27].

Appendix D. Proof of Proposition 5.1. Consider problem (5.5) *without* the KL constraint, and with the values $w(k)$ arranged in increasing order (see Figure D.1):

$$\varphi(\mu) \doteq \max_{\pi} \sum_{k=1}^{T} \pi_k |w(k) - \mu| \quad \text{subject to} \quad \begin{array}{r} \mathbf{w}^\top \pi = \mu, \\ \mathbf{1}^\top \pi = 1, \\ \pi \geq 0. \end{array} \tag{D.1}$$

It is immediate by inspection that (5.7) is a feasible solution for (D.1). We next show that this solution is actually optimal.

Let $d(k) \doteq |w(k) - \mu|$. The Lagrangian for problem (D.1) is

$$\mathcal{L}(\pi, \lambda, \nu) = -\sum_{k=1}^{T} \pi_k d(k) - \sum_{k=1}^{T} \lambda_k \pi_k + \nu_1 \left(\sum_{k=1}^{T} \pi_k w(k) - \mu \right) + \nu_2 \left(\sum_{k=1}^{T} \pi_k - 1 \right).$$

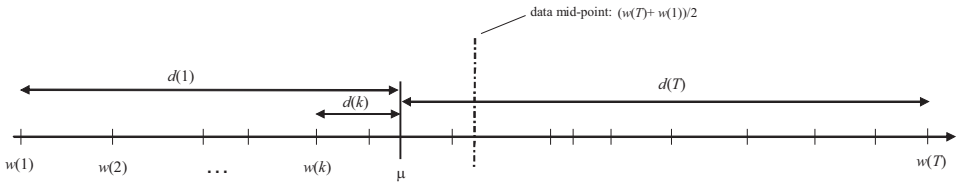


FIG. D.1. *Data arranged in increasing order. The $d(k)$ are the distances from the mean: $d(k) \doteq |w(k) - \mu|$.*

For a feasible π to be optimal, the Karush–Kuhn–Tucker (KKT) condition

$$\nabla_\pi \mathcal{L}(\pi, \lambda, \nu) = -\begin{bmatrix} d(1) \\ \vdots \\ d(T) \end{bmatrix} - \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_T \end{bmatrix} + \nu_1 \begin{bmatrix} w(1) \\ \vdots \\ w(T) \end{bmatrix} + \nu_2 \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = 0 \tag{D.2}$$

must hold for some ν_1, ν_2 and $\lambda_k \geq 0$, along with the complementary slackness condition

$$\lambda_k \pi_k = 0, \quad k = 1, \ldots, T. \tag{D.3}$$

We now show that for the solution in (5.7) we can find ν_1, ν_2 and $\lambda_k \geq 0$ such that (D.2), (D.3) hold. From (D.2) we have

$$\lambda_k = \nu_1 w(k) + \nu_2 - d(k), \quad k = 1, \ldots, T. \tag{D.4}$$

Since $\pi_1 > 0, \pi_T > 0$, (D.3) implies $\lambda_1 = \lambda_T = 0$, and hence

$$\begin{aligned} \nu_1 w(1) + \nu_2 - d(1) &= 0, \\ \nu_1 w(T) + \nu_2 - d(T) &= 0, \end{aligned}$$

from which we obtain

$$\nu_1 = \frac{d(T) - d(1)}{w(T) - w(1)}, \quad \nu_2 = \frac{d(1)[w(T) - w(1)] - w(1)[d(T) - d(1)]}{w(T) - w(1)}. \tag{D.5}$$

Substituting into (D.4), for $k = 2, \ldots, T-1$, we get

$$\lambda_k = \frac{[d(T) - d(1)][w(k) - w(1)] + [w(T) - w(1)][d(1) - d(k)]}{w(T) - w(1)}. \tag{D.6}$$

We now verify that these λ_k are nonnegative. Define $K_- \doteq \{k : w(k) \leq \mu\}$ and $K_+ \doteq \{k : w(k) > \mu\}$. For $k \in K_-$ we have that (see Figure D.1)

$$w(k) - w(1) = d(1) - d(k); \quad w(T) - w(1) = d(1) + d(T),$$

which, once substituted in (D.6), give

$$\begin{aligned} \lambda_k &= \frac{[d(T) - d(1)][d(1) - d(k)] + [d(1) + d(T)][d(1) - d(k)]}{w(T) - w(1)} \\ &= \frac{2d(T)[d(1) - d(k)]}{w(T) - w(1)} \geq 0, \quad k \in K_-. \end{aligned}$$

Similarly, for $k \in K_+$ we have that

$$w(k) - w(1) = d(1) + d(k); \quad w(T) - w(1) = d(1) + d(T),$$

which, once substituted in (D.6), give

$$\begin{aligned}\lambda_k &= \frac{[d(T) - d(1)][d(1) + d(k)] + [d(1) + d(T)][d(1) - d(k)]}{w(T) - w(1)} \\ &= \frac{2d(1)[d(T) - d(k)]}{w(T) - w(1)} \geq 0, \quad k \in K_+.\end{aligned}$$

Overall, we have that the primal feasible solution (5.7), together with the dual feasible variables (D.5), (D.6), satisfies the KKT conditions (D.2), (D.3), and hence (5.7) is optimal for problem (D.1). If this solution satisfies the constraint $\mathrm{KL}(\pi, \eta) \leq d$, then clearly the solution is also optimal for the original problem (5.5).

Substituting (5.7) into the objective (D.1), we easily obtain the optimal value function in (5.8), which concludes the proof. □

Acknowledgments. I sincerely thank Fabrizio Dabbene, Constantino Lagoa, Arkadi Nemirovski, and Andrzej Ruszczyński for their useful discussions and correspondence. A special thanks goes to the Associate Editor and to the anonymous reviewers for their precious comments and suggestions.

REFERENCES

[1] S. M. Ali and S. D. Silvey, *A general class of coefficients of divergence of one distribution from another*, J. Royal Statist. Soc., 28 (1966), pp. 131–142.
[2] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath, *Coherent measures of risk*, Math. Finance, 9 (1999), pp. 203–228.
[3] V. S. Bawa, S. J. Brown, and R. W. Klein, *Estimation Risk and Optimal Portfolio Choice*, North-Holland, Amsterdam, 1979.
[4] A. Ben-Tal and A. Nemirovskii, *Robust convex optimization*, Math. Oper. Res., 23 (1998), pp. 769–805.
[5] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.
[6] M. Broadie, *Computing efficient frontiers using estimated parameters*, Ann. Oper. Res., 45 (1993), pp. 21–58.
[7] Z. Chen and L. Epstein, *Ambiguity, risk, and asset returns in continuous time*, Econometrica, 70 (2002), pp. 1403–1443.
[8] V. K. Chopra and W. T. Ziemba, *The effect of errors in means, variances and covariances on optimal portfolio choice*, J. Portfolio Management, 19 (2) (1993), pp. 6–11.
[9] D. den Hertog, F. Jarre, C. Roos, and T. Terlaky, *A sufficient condition for self-concordance, with application to some classes of structured convex programming problems*, Math. Programming, 69 (1995), pp. 75–88.
[10] R. M. Dudley, *Distance of probability measures and random variables*, Ann. Math. Statist., 39 (1968), pp. 1563–1572.
[11] J. Dupačová, *The minimax approach to stochastic program and illustrative application*, Stochastics, 20 (1987), pp. 73–88.
[12] J. Dupačová, *Stochastic Programming: Minimax Approach*, in Encyclopedia of Optimization, Vol. 5, Kluwer Academic, Dordrecht, The Netherlands, 2001, pp. 327–330.
[13] L. El Ghaoui, M. Oks, and F. Oustry, *Worst-case value-at-risk and robust portfolio optimization: A conic programming approach*, Oper. Res., 51 (2003), pp. 543–556.
[14] L. El Ghaoui, F. Oustry, and H. Lebret, *Robust solutions to uncertain semidefinite programs*, SIAM J. Optim., 9 (1998), pp. 33–52.
[15] E. Erdoğan and G. Iyengar, *Ambiguous chance constrained problems and robust optimization*, Math. Program. Ser. B, 107 (2006), pp. 37–61.
[16] J.-L. Goffin, Z.-Q. Luo, and Y. Ye, *Complexity analysis of an interior cutting plane method for convex feasibility problems*, SIAM J. Optim., 6 (1996), pp. 638–652.

[17] J.-L. Goffin and J.-P. Vial, *Convex nondifferentiable optimization: A survey focused on the analytic center cutting plane method*, Optim. Methods Softw., 17 (2002), pp. 805–867.
[18] D. Goldfarb and G. Iyengar, *Robust portfolio selection problems*, Math. Oper. Res., 28 (2003), pp. 1–38.
[19] P. Hall, *On Kullback-Leibler loss and density estimation*, Ann. Statist., 15 (1987), pp. 1491–1519.
[20] M. Henry, *Generalized Entropy Measure of Ambiguity and Its Estimation*, preprint, Columbia University, New York, 2005.
[21] H. Konno and H. Yamazaki, *Mean absolute deviation portfolio optimization model and its application to Tokio stock market*, Management Sci., 37 (1991), pp. 519–531.
[22] S. Kullback, *Information Theory and Statistics*, Wiley, New York, 1959.
[23] M. S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret, *Applications of second-order cone programming*, Linear Algebra Appl., 284 (1998), pp. 193–228.
[24] P. J. Maenhout, *Robust portfolio rules and asset pricing*, Rev. Financial Stud., 17 (2004), pp. 951–983.
[25] H. M. Markowitz, *Portfolio selection*, J. Finance, 7 (1952), pp. 77–91.
[26] R. O. Michaud, *The Markowitz optimization enigma: Is optimized optimal?*, Financial Analysts J., 45 (1989), pp. 31–42.
[27] Yu. Nesterov, *Smooth Minimization of Nonsmooth Functions*, CORE Tech. Report, 2003.
[28] Yu. Nesterov and A. S. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM, Philadelphia, 1994.
[29] O. Péton, *The Homogeneous Analytic Center Cutting Plane Method*, Ph.D. Thesis, Université de Genève, 2002.
[30] A. Ruszczyński and A. Shapiro, *Optimization of risk measures*, in Probabilistic and Randomized Methods for Design under Uncertainty, G. Calafiore and F. Dabbene, eds., Springer-Verlag, London, 2006, pp. 119–157.
[31] A. Ruszczynski and R. J. Vanderbei, *Frontiers of stochastically nondominated portfolios*, Econometrica, 71 (2003), pp. 1287–1297.
[32] A. Shapiro and A. J. Kleywegt, *Minimax analysis of stochastic problems*, Optim. Methods Softw., 17 (2002), pp. 523–542.
[33] M. J. Todd, *Semidefinite optimization*, Acta Numer., 10 (2001), pp. 515–560.
[34] R. H. Tütüncü and M. Koenig, *Robust asset allocation*, Ann. Oper. Res., 132 (2004), pp. 157–187.
[35] L. Vandenberghe and S. Boyd, *Semidefinite programming*, SIAM Rev., 38 (1996), pp. 49–95.
[36] J. Žáčková, *On minimax solutions of stochastic linear programming problems*, Čas. Pěst. Mat., 91 (1966), pp. 423–430.
[37] Z. Wang, *A shrinkage approach to model uncertainty and asset allocation*, Rev. Financial Stud., 18 (2005), pp. 673–705.

SIAM J. OPTIM.
Vol. 18, No. 3, pp. 878–894

ON THE REGULARIZATION OF SLIDING MODES*

LAURA LEVAGGI† AND SILVIA VILLA†

Abstract. Approximability of sliding motions for control systems governed by nonlinear finite-dimensional differential equations is considered. This regularity property is shown to be equivalent to Tikhonov well-posedness of a related minimization problem in the context of relaxed controls. This allows the derivation of a general approximability result, which in the autonomous case has an easy to verify geometrical formulation. In the second part of the paper, we consider nonapproximable sliding mode control systems. In the flavor of regularization of ill-posed problems, we propose a method of selection of well-behaved approximating trajectories converging to a prescribed ideal sliding.

Key words. sliding mode control, approximability, Tikhonov well-posedness, regularization

AMS subject classifications. 93B12, 49K40, 49J45

DOI. 10.1137/060657157

1. Introduction. Sliding mode control methods aim at fulfilling the control objective by constraining the system motion on a prescribed, suitably chosen manifold S (reference books on theory and applications are [12, 16]). In real-life applications, however, for various reasons the reaching of the ideal sliding (i.e., the evolution on S) is prevented. In fact, the required finite-time global attractivity of S forces the use of state-discontinuous feedback controllers. This is impossible in practice, both because nonidealities in the actuators such as small delays, hysteresis, and so on have to be taken into account and because discontinuity induces infinitely fast switching around the sliding surface (chattering), which is potentially damaging for mechanical components. Therefore the sliding condition can be only approximately satisfied, giving rise to the so-called real sliding motions. Hence, a key issue in sliding mode control theory is to establish conditions under which it is assured that real states enjoy the same dynamical properties of the ideal sliding. In the literature this topic goes under the broad definition of approximability [6, 7, 8, 9, 16, 17]. More precisely, a first result about the convergence of real sliding motions to the evolution obtained with the equivalent control method for affine control systems can be found in section 2.3 of [16]. A general mathematical formulation of approximability, requiring existence and uniqueness of the equivalent control, was then given in [8]. The authors consider nonlinear control systems, and approximability is proved for classes of systems satisfying particular structural properties. In [9] a new definition of approximability, called generalized approximability, is introduced: the new concept does not require the existence of the equivalent control. Criteria guaranteeing generalized approximability are obtained, extending results in [8]. Approximability and related regularity properties also have been investigated for second and higher order sliding modes in [6, 7]. A new and more appropriate definition of real states is then given in [17]: the class of approximating trajectories is enlarged to comprise both classical solutions

*Received by the editors April 11, 2006; accepted for publication (in revised form) February 5, 2007; published electronically October 4, 2007. This research was partially supported by MIUR cofinanced project "Control, optimization and stability of non-linear systems: geometrical and analytical methods."

http://www.siam.org/journals/siopt/18-3/65715.html

†Dipartimento di Matematica, Università di Genova, Via Dodecaneso 35, 16146 Genova, Italy (levaggi@dima.unige.it, villa@dima.unige.it).

corresponding to continuous controls and evolutions generated by discontinuous feedback controls in the Filippov sense. The corresponding definition of approximability is therefore strengthened. Moreover, as a main result, it is shown that approximability is equivalent to Tikhonov well-posedness of a suitably defined minimization problem. New approximability criteria are then obtained as a consequence of known results characterizing well-posedness (for an introduction about this theory, see, e.g., [11]).

Nonapproximability is therefore a consequence of ill-posedness of a minimization problem on the set of trajectories of the control system. This in general means that there exist sequences of real sliding motions that either do not converge to a system motion or can approach different sliding trajectories. Is it then possible to use a regularization procedure to suitably choose approximating sequences and thus partially restore well-posedness? In this paper we are going to give a possible answer to this question. A classical regularization method is Tikhonov regularization, i.e., to consider approximate solutions of perturbed minimization problems obtained by adding a small uniformly convex term to the objective function (see Chapter I, section 7 of [11]). However, a straightforward application of this procedure is not possible in the setting of [17]. For example, it is not clear whether the set of trajectories $W(x_0)$, issued from an arbitrary point x_0, resulting from that choice of admissible controls is closed in the norm of uniform convergence. The key result needed to prove this property is contained in Theorem 2.1. There we show that, whenever the set of control values U is compact, $W(x_0)$ can be characterized as the solution set of a differential inclusion, which in turn coincides with the set of trajectories resulting from the application of relaxed controls (see Corollary 2.3). Even if the use of relaxed controls is a novelty in the context of sliding mode control theory, this result shows the existence of an underlying natural connection. In our paper we take advantage of this characterization in order to both show the compactness of $W(x_0)$ and prove in a more direct way the equivalence in [17] between approximability and well-posedness. The use of relaxed controls in fact avoids distinctions between Carathéodory and Filippov solutions, since these are now unified under a global framework. Moreover, by the compactness of $W(x_0)$, approximability is equivalent to the existence of a unique sliding motion. This, combined with the characterization of $W(x_0)$, allows us to obtain as a main result a complete geometrical characterization of approximability for regular autonomous sliding control systems. In these cases knowledge of both the velocity vector field and the tangent cone to the sliding surface is sufficient to establish whether approximability holds. Therefore, in contrast to previous results, a priori knowledge of the solutions is not required, and we obtain well-posedness without convexity assumptions. The compactness of $W(x_0)$ is furthermore exploited to devise a regularization strategy, allowing more flexibility in the design of the perturbations of the minimization problem. Theorem 4.1 gives a general, new regularization result, essentially different from Tikhonov's, inspired by variational convergences. The idea is to construct a sequence of perturbations of the given problem in such a way that the asymptotically minimizing sequences do converge to a well-defined sliding mode. A general method for choosing these regularizing functionals is also given when an equivalent control is known.

The paper is organized as follows: in section 2 we introduce the general setting of the problem with the standing regularity hypotheses. Next, we study the set of admissible trajectories $W(x_0)$, discussing in detail the choice of the class of admissible controls $\mathcal{U}$. We show in particular that, whenever U is compact and $\mathcal{U}$ is chosen as in [17], the set $W(x_0)$ coincides with the evolutions of the controlled system through the application of relaxed controls. This is contained in the main result of the sec-

tion, Theorem 2.1 along with Corollary 2.3. Section 3 is devoted to the relationship between approximability and well-posedness. The equivalence in our setting is proved in Theorem 3.1 using relaxed controls. Then, in Theorem 3.4, we give a geometrical characterization of approximability for regular autonomous systems, extending previous results. Finally, a class of control systems is presented for which approximability holds under general regularity assumptions. Section 4 is dedicated to the regularization of ill-posed sliding mode control systems. The classical nonapproximable model problem proposed by Izosimov is presented under a new perspective, showing in detail its main features. Then we give the main regularization result, Theorem 4.1, and discuss its applications.

2. The sliding mode control system. In this section we introduce the class of control systems and sliding manifolds we are going to consider, along with the family $\mathcal{U}$ of admissible controls of [17]. $\mathcal{U}$ contains both Carathéodory and state-discontinuous feedbacks, and the corresponding set $W(x_0)$ of admissible trajectories for an arbitrary starting point x_0 is made up of Carathéodory and Filippov solutions of closed loop equations. Real sliding motions are then defined as evolutions generated by admissible feedbacks, which fulfill only approximately the sliding constraint. Since in what follows we will be interested in examining their limit behavior, a better characterization of the sets $W(x_0)$ is very helpful. In Theorem 2.1 we show that

$$W(x_0) = \{x \in AC(0,T;\mathbb{R}^N) : \dot{x}(t) \in \operatorname{co} f(t,x(t),U), \text{ for a.e. } t,\ x(0) = x_0\}$$

for any x_0, whenever the set of control values U is compact. The differential inclusion in the above identity is related to relaxation techniques. In particular we can read this equality as a dynamic equivalence between the application (in the Filippov sense) of discontinuous feedbacks and relaxed controls.

Through the obtained characterization we are also able to directly prove the compactness of $W(x_0)$ in the supremum norm, under the regularity hypotheses of our setting. All subsequent results, not least the regularization of nonapproximable sliding mode control systems, will depend upon this key property.

2.1. General setting and regularity assumptions. We consider the following control system:

$$\begin{cases} \dot{x} = f(t,x,u), & x \in \mathbb{R}^N,\ u \in U,\ t \in [0,T], \\ x(0) = x_0 \end{cases} \tag{1}$$

with

$$f : [0,T] \times \mathbb{R}^N \times U \to \mathbb{R}^N$$

a Carathéodory function, $U \subset \mathbb{R}^M$ compact, and T fixed. Throughout the paper the following assumptions on f will be assumed to hold:

(F1) There exist $A, B \in L^1(0,T)$ such that

$$|f(t,x,u)| \le A(t)|x| + B(t), \qquad \text{for a.e. } t, \quad \forall x \in \mathbb{R}^N,\ \forall u \in U. \tag{2}$$

(F2) For any compact subset Z of $\mathbb{R}^N$ there exists $C \in L^1(0,T)$ such that

$$|f(t,x_1,u) - f(t,x_2,u)| \le C(t)|x_1 - x_2|, \quad \text{for a.e. } t \in [0,T], \quad \forall x_1, x_2 \in Z,\ \forall u \in U. \tag{3}$$

As for the sliding manifold, we suppose that, for any t, it is defined as the set of zeros of a continuous function

$$s : [0,T] \times \mathbb{R}^N \to \mathbb{R}^M;$$

i.e., $S(t) = \{x \in \mathbb{R}^N : s(t,x) = 0\}$.

2.2. The choice of admissible controls. In [17] the class $\mathcal{U}$ of admissible feedback control laws $u(t,x)$ is a subset of $L \otimes B$-measurable functions. Solutions of the resulting closed loop equation are intended in either the Carathéodory or Filippov sense, so that the set of trajectories of (1) corresponding to these admissible controls is

(4)
$$W(x_0) = \{x \in AC(0,T;\mathbb{R}^N) : \exists\, u \in \mathcal{U} \text{ s.t. } \dot{x} = f(t,x,u(t,x)) \text{ for a.e. } t,\ x(0) = x_0, \text{ or } \dot{x} \in F_u(t,x) \text{ for a.e. } t,\ x(0) = x_0\}.$$

Recall that for a feedback law $u(t,x)$ the Filippov multifunction is defined as

$$F_u(t,x) := \bigcap_{\varepsilon>0} \bigcap_{|N|=0} \overline{\text{co}}\, f(t, B(x,\epsilon) \setminus N, u(t, B(x,\epsilon) \setminus N)), \tag{5}$$

where $B(x,\epsilon)$ is the ball of center x and radius ϵ in $\mathbb{R}^N$ and $|N|$ is the Lebesgue measure of the set N. More precisely, in [17],

$$\mathcal{U} = \{u : [0,T]\times\mathbb{R}^N \to U,\ L\otimes B\text{-meas.},\ \|u(\cdot,x(\cdot))\|_\infty < \infty\ \forall x \in W(x_0),\ \forall x_0 \in \mathbb{R}^N\}.$$

If U is compact, the condition $\|u(\cdot,x(\cdot))\|_\infty < \infty$ is unnecessary, and it is possible to give a simpler characterization of $W(x_0)$.

THEOREM 2.1. *Let the hypotheses of the general setting be satisfied. Then*

$$W(x_0) = \{x \in AC(0,T;\mathbb{R}^N) : \dot{x}(t) \in \text{co}\, f(t,x(t),U),\ \textit{for a.e. } t,\ x(0) = x_0\}. \tag{6}$$

Proof. Let $z \in W(x_0)$. If z is a Carathéodory solution of (1), it trivially belongs to the right-hand side set of (6). It is also easy to see that for any admissible feedback u the Filippov multifunction F_u in (5) satisfies $F_u(t,x) \subseteq \overline{\text{co}}\, f(t,x,U)$ because of the continuity of f in the state variable. Since U is compact $\overline{\text{co}}\, f(t,x,U) = \text{co}\, f(t,x,U)$, therefore $W(x_0)$ is included in $\{x \in AC(0,T;\mathbb{R}^N) : \dot{x}(t) \in \text{co}\, f(t,x(t),U)$, for a.e. t, $x(0) = x_0\}$.

To prove that the converse is true, let $x : [0,T] \to \mathbb{R}^N$ be absolutely continuous and such that

$$\dot{x}(t) \in \text{co}\, f(t,x(t),U), \qquad \text{for a.e. } t \in [0,T], \quad x(0) = x_0.$$

Then, by Carathéodory's convexity theorem, we can write

$$\dot{x}(t) = \sum_{i=1}^{N+1} \tilde{\lambda}_i(t) f(t,x(t),\tilde{u}_i(t)), \qquad t \notin E,$$

for some E with $|E| = 0$ and $(\tilde{\lambda}(t),\tilde{u}(t)) \in [0,1]^{N+1} \times U^{N+1} =: X$. Let

$$\Gamma(t) = \begin{cases} \left\{(\lambda,u) \in X : \dot{x}(t) = \displaystyle\sum_{i=1}^{N+1} \lambda_i f(t,x(t),u_i)\right\}, & t \notin E, \\ \{(\lambda_0,u_0)\}, & t \in E, \end{cases}$$

with $(\lambda_0, u_0) \in [0,1] \times U$ arbitrarily fixed. Then obviously Γ has nonempty values. Moreover, by Corollary 1Q in [15] Γ is a closed-valued, measurable multifunction, since $\dot{x}$ is measurable, $\lambda \mapsto \sum_{i=1}^{N+1} \lambda_i f(t, x(t), u_i)$ is continuous, and f is Carathéodory. Therefore by Corollary 1C in [15] there exists a measurable selection $(\lambda(t), u(t)) \in \Gamma(t)$ for all t. Now we use the selected $u(t)$ in order to define a feedback control law $\hat{u}$ such that $x(\cdot)$ is a Filippov solution of the corresponding closed loop equation. The required feedback $\hat{u}$ has to be discontinuous at every point of the trajectory $x(t)$ and be such that in any neighborhood of $x(t)$ all values $u_1(t), \dots, u_{N+1}(t)$ are taken on. To this end we define the following partition of $[0,T] \times \mathbb{R}^N$:

$$\begin{aligned} A_1 &= \{(t,x) : x_1 < x_1(t), \dots, x_N < x_N(t)\}, \\ A_i &= \{(t,x) : x_k > x_k(t) \text{ for } k < i,\ x_r < x_r(t) \text{ for } r \geq i\}, \quad i = 2, \dots, N, \\ A_{N+1} &= \{(t,x) : x_1 > x_1(t), \dots, x_N > x_N(t)\}, \\ A_0 &= \complement \left(\bigcup_{i=1}^{N+1} A_i \right). \end{aligned}$$

Note that for arbitrary t, $\epsilon > 0$, $Z \subset \mathbb{R}^N$, with $|Z| = 0$, the set $B(x(t), \epsilon) \setminus Z$ has nonvoid intersection with every A_i, $i = 1, \dots, N+1$. Now, if $u_0(t) = u_0$, it is straightforward to see that the control law $\hat{u}$,

$$\hat{u}(t,x) = \sum_{i=0}^{N+1} u_i(t)\, \chi_{A_i}(t,x),$$

where χ_{A_i} is the characteristic function of the set A_i, has the required features. Thanks to the properties of the chosen partition, $x(\cdot)$ is a solution of the Filippov differential inclusion (see (5))

$$\dot{x}(t) \in F_{\hat{u}}(t, x(t)), \text{ for a.e. } t, \qquad x(0) = x_0.$$

It remains to prove that $\hat{u}$ is $L \otimes B$-measurable, i.e., that all the sets A_i, $i = 0, \dots, N+1$, are measurable. Since the functions $g_j(t,x) = x_j - x_j(t)$ for $j = 1, \dots, N$ are continuous, the sets $g_j^{-1}(0, +\infty)$ and $g_j^{-1}(-\infty, 0)$ are Borel sets for any j. Therefore the sets A_i are a finite intersection of Borel sets and hence measurable. ∎

The next lemma establishes the compactness of $W(x_0)$, which will be extensively used in what follows.

LEMMA 2.2. *Let $x_0 \in \mathbb{R}^N$ and z_n be any sequence in $W(x_0)$. Then z_n is equibounded and equicontinuous. Moreover, the set $W(x_0)$ is closed and thus compact as a subset of $C^0(0,T;\mathbb{R}^N)$ endowed with the norm of uniform convergence.*

Proof. The result depends upon the regularity properties of f and the compactness of U. Since f is Carathéodory, the set-valued map $F_1(t,x) = f(t,x,U)$ has compact values, is upper semicontinuous in x for a.e. t (see, e.g., Proposition 1.4.14 in [5]), and is measurable in t for any x. Thus $F(t,x) = \operatorname{co} F_1(t,x)$ enjoys the same properties, plus convexity of its values. Moreover, the Carathéodory convexity theorem and (2) yield

$$\begin{aligned} \|F(t,x)\| &= \sup \left\{ \left| \sum_{i=0}^{N} \alpha_i f(t,x,u_i) \right| : \alpha_i \in [0,1],\ \sum_{i=0}^{N} \alpha_1 = 1,\ u_i \in U \right\} \\ &\leq C(t)(1 + |x|) \end{aligned}$$

for some $C \in L^1(0,T)$. Therefore the hypotheses of Theorem 7.1 in [10] are satisfied, and the result follows. □

We have thus shown that if U is compact, the set in (4) is in fact the set of solutions of the differential inclusion $\dot{x}(t) \in \operatorname{co} f(t, x(t), U)$. It is a classical result (see, e.g., Theorem 13.4.1 in [13]) that the latter corresponds to the set of trajectories of (1) under the action of relaxed control laws [1, 2, 13]. These are measure-valued controls defined as functions $\mu(\cdot)$:

$$\mu : [0,T] \to \Sigma(U), \quad \mu \in L^\infty_w(0,T;\mathcal{M}(U)) = L^1(0,T;C(U))^*,$$

where $C(U)$ is the set of continuous functions on U, $\mathcal{M}(U) = C(U)^*$ is the space of Borel measures on U, and $\Sigma(U) = \{\mu \in \mathcal{M}(U) : \mu \geq 0,\ \mu(U) = 1\}$ is the set of Borel probability measures on U. Solutions of (1) under the application of a relaxed control law μ are absolutely continuous functions $x : [0,T] \to \mathbb{R}^N$ such that

$$x(t) = x_0 + \int_0^t \left(\int_U f(t, x(t), u)\, \mathrm{d}\mu_t \right) dt, \tag{7}$$

where μ_t denotes the measure $\mu(t) \in \Sigma(U)$. Therefore, we have the following corollary.

COROLLARY 2.3. *Let assumptions* (F1)–(F2) *on the function f be satisfied and U be compact. Call $R(x_0)$ the set of solutions of* (7) *as μ varies in the set of relaxed controls. Then*

$$\begin{aligned} W(x_0) &= \{x \in AC :\ \dot{x}(t) \in \operatorname{co} f(t, x(t), U),\ \textit{for a.e. } t,\ x(0) = x_0\} \\ &= R(x_0). \end{aligned} \tag{8}$$

This result gives us the possibility of taking advantage of the mathematical structure offered by relaxed controls while retaining $\mathcal{U}$ as the class of admissible controls, which is a more natural choice from a practical point of view.

3. Approximability and well-posedness. In this section we study the relationship between approximability of the sliding mode and Tikhonov well-posedness of a suitable minimization problem. After giving the definitions of sliding states and approximability for sliding motions, we introduce the concept of Tikhonov well-posedness. As in [17] we then prove that these regularity notions are equivalent. Moreover, we show that Tikhonov well-posedness (and thus approximability) has an easy characterization in our setting. Then we specialize our results in the autonomous case: for these systems, under some reasonable regularity assumptions, we are in fact able to give a purely geometric necessary and sufficient condition for approximability of the sliding mode control system. In the last part of the section we present a class of systems and surfaces for which sliding is well-posed and to which approximability criteria in [17] are not, in general, applicable.

3.1. Sliding modes approximability and Tikhonov well-posedness. Let us now consider the control system (1) in the general setting described in section 2.1. For any $x_0 \in \mathbb{R}^N$ such that $s(0, x_0) = 0$, we say that $y \in W(x_0)$ is a *sliding state* issued from x_0 if $s(t, y(t)) = 0$ for all $t \in [0,T]$. In the terminology of [3, 4, 5], y is a solution of the differential inclusion in (8), viable on the (time-varying) sliding manifold.

As in [17] we introduce the following *approximability property* for sliding modes:

(A) for any $x_0 \in S(0)$ there exists a unique sliding state y issued from x_0, and for any sequences $x_n \to x_0$, $y_n(\cdot) \in W(x_n)$ such that $\|s(\cdot, y_n(\cdot))\|_\infty \to 0$, one has $\|y_n - y\|_\infty \to 0$.

Property (A) is related to the Tikhonov well-posedness with value zero [11, Chap. 1] of the following optimization problem:

$$\min_{x\in W(x_0)} I(x), \qquad I(x)=\int_0^T |s(t,x(t))|\,dt,$$

where $|\cdot|$ is any norm in $\mathbb{R}^N$. In what follows we will refer to the above problem by writing $(W(x_0),I)$. The well-posedness requirement for $(W(x_0),I)$ consists in the following condition:

(TWP) there exists a unique minimizer $y\in W(x_0)$ for I, $I(y)=0$, and any minimizing sequence $z_n\in W(x_0)$ uniformly converges to y; i.e., $I(z_n)\to 0$ implies $\|z_n-y\|_\infty\to 0$.

Using Lemma 2.2 and the characterization (8) it is possible to prove the equivalence of properties (A) and (TWP) as in Theorem 3.1 in [17] under a unified framework, avoiding the distinction between different concepts of solution of (1).

THEOREM 3.1. *Let the hypotheses of the general setting be satisfied and x_0 be any vector such that $s(0,x_0)=0$. Then $(W(x_0),I)$ is Tikhonov well-posed with value zero for any such x_0 if and only if the approximability property* (A) *is fulfilled.*

Proof. Let x_0 be fixed; suppose that the problem $(W(x_0),I)$ satisfies property (TWP). Let x_n and y_n be as in condition (A). Then by Corollary 2.3 there exist relaxed controls μ^n such that

$$y_n(t)=x_n+\int_0^t\left(\int_U f(s,y_n(s),u)\,\mathrm{d}\mu_s^n\right)\mathrm{d}s.$$

Then calling z_n the solution of the integral equation

$$z_n(t)=x_0+\int_0^t\left(\int_U f(s,z_n(s),u)\,\mathrm{d}\mu_s^n\right)\mathrm{d}s,$$

using (3) and Lemma 2.2, and recalling that μ_s^n is a probability measure for any n and s, we obtain

$$\begin{aligned}|y_n(t)-z_n(t)| &\le |x_n-x_0|+\int_0^t \sup_{u\in U}|f(s,y_n(s),u)-f(s,z_n(s),u)|\,\mathrm{d}s\\ &\le |x_n-x_0|+\int_0^t C(s)|y_n(s)-z_n(s)|\,\mathrm{d}s,\end{aligned}$$

and by Gronwall's lemma $\|y_n-z_n\|_\infty\to 0$. Thus $\|s(\cdot,z_n)\|_\infty\to 0$, and a fortiori z_n is a minimizing sequence for $(W(x_0),I)$. Tikhonov well-posedness then implies the convergence of z_n, and thus of y_n, to the unique sliding state y as desired.

Suppose now that the approximability property (A) holds. To show the well-posedness of $(W(x_0),I)$, let z_n be a minimizing sequence. Then by Lemma 2.2 any of its subsequences admits a further subsequence which is uniformly convergent to some $z\in W(x_0)$. Also, since $s(\cdot,z_n(\cdot))\to 0$ in $L^1(0,T)$, eventually passing to a further subsequence (we do not relabel for simplicity), we have $s(t,z_n(t))\to 0$ a.e. on $[0,T]$. By continuity of s and uniform convergence,

$$s(t,z_n(t))\to 0, \quad s(t,z_n(t))\to s(t,z(t)) \quad \text{uniformly in } t.$$

Approximability then implies that z coincides with the unique sliding state y. Since this limit is independent of the chosen subsequence, we get the convergence of z_n to y as desired. □

Another consequence of Lemma 2.2 is the following characterization of Tikhonov well-posedness of $(W(x_0), I)$ and thus of the approximability of the associated sliding mode control system.

COROLLARY 3.2. *Property* (TWP) *is satisfied if and only if problem* $(W(x_0), I)$ *admits a unique minimizer* y *with* $I(y) = 0$.

We now show a geometric characterization of well-posedness in the above sense for a general class of sliding mode control systems. Before stating this result we recall the following lemma.

LEMMA 3.3 (Lemma 4 in [14]). *Let* $A \subset \mathbb{R}^N$ *be an open set, and let* G *and* H *be two set-valued maps from* A *to* $\mathbb{R}^N$. *Assume that* G *and* H *have nonempty convex closed values, that* G *is continuous, and that* H *is lower semicontinuous. If*

$$G(x) \cap H(x) \neq \varnothing \qquad \forall x \in A,$$

then there exists a continuous function $f : A \to \mathbb{R}^N$ *with*

$$f(x) \in G(x) \cap H(x) \qquad \forall x \in A.$$

Also, recall that for a subset S of $\mathbb{R}^N$ the *Bouligand tangent cone* $T_S(x)$ is defined as

$$T_S(x) = \left\{ v \in \mathbb{R}^N : \liminf_{h \to 0^+} \frac{d(x + hv, S)}{h} = 0 \right\}$$

and that S is *sleek* if the multivalued map $x \mapsto T_S(x)$ is lower semicontinuous (see, e.g., section 4.1.4 in [5] for the definition and properties of sleek subsets).

THEOREM 3.4. *Let us consider a general autonomous control system of the form*

$$\begin{cases} \dot{x} = f(x, u), \\ x(0) = x_0 \end{cases} \tag{9}$$

with f *satisfying conditions* (F1)–(F2) *and a time-independent sliding manifold* S *which is a sleek subset of* $\mathbb{R}^N$. *Then for any* $x_0 \in S$ *a sliding mode on* S *in* $W(x_0)$ *exists if and only if*

$$\operatorname{co} f(x, U) \cap T_S(x) \neq \varnothing \qquad \forall x \in S.$$

Moreover, let the above condition be satisfied; then the corresponding sliding motion is unique for any $x_0 \in S$ *if and only if*

$$\operatorname{co} f(x, U) \cap T_S(x) = \{\tau(x)\} \qquad \forall x \in S \tag{10}$$

with $\tau : S \to \mathbb{R}^N$ *such that any Cauchy problem associated with the differential equation* $\dot{x} = \tau(x)$ *admits a unique solution on* $[0, T]$.

Proof. By (6), the set of sliding modes of (9) on S will be given by solutions of $\dot{x}(t) \in F(x(t)) := \operatorname{co} f(x(t), U)$ such that $x(t) \in S$ for all t. Therefore by Lemma 1.1.4 in [3] any sliding mode $y(\cdot)$ must satisfy

$$\dot{y}(t) \in F(y(t)) \cap T_S(y(t)), \quad \text{for a.e. } t. \tag{11}$$

Also, by standard viability results (see, e.g., Proposition 4.2.1 and Theorem 4.2.1 in [4]) a sliding mode exists if and only if the intersection $F(x) \cap T_S(x)$ is nonvoid for any $x \in S$.

Moreover, if (10) is satisfied, from (11) the admissible sliding motions are the solutions of $\dot{x} = \tau(x)$. Thus uniqueness depends on the regularity of τ.

Suppose now that a sliding mode exists, but the intersection in (10) is not single-valued. By assumption (F2) we have Lipschitz continuity for F and thus continuity. Let $x_0 \in S$ and $y_0 \in F(x_0) \cap T_S(x_0)$; now define $H(x_0) = \{y_0\}$ and $H(x) = T_S(x)$ otherwise. Then H is lower semicontinuous, and by Lemma 3.3 there exists a continuous selection for the map $x \mapsto F(x) \cap H(x)$. Therefore for any $y_0 \in F(x_0) \cap T_S(x_0)$ we can construct a selection g of $F(\cdot) \cap T_S(\cdot)$ such that $g(x_0) = y_0$. By continuity, this shows that if the intersection between the velocity set $F(x)$ and the tangent cone $T_S(x)$ is not single-valued on S, different sliding modes exist in $W(x_0)$, since different selections can be constructed. □

As a consequence, by Theorem 3.1 and Corollary 3.2, we have the following corollary.

COROLLARY 3.5. *A sliding mode control problem with time-independent sleek sliding manifold for an autonomous nonlinear system is approximable for every $x_0 \in S$ if and only if condition* (10) *holds.*

Let us show a very simple and classical example.

Let $f(x,u) = Ax + Bu$ and $S = \{x \in \mathbb{R}^N : Cx = 0\}$ with A, B, and C matrices of the right dimensions. Then $T_S(x) = S$ for all $x \in S$, and condition (10) is satisfied if and only if the matrix CB is invertible. In this case there exists a uniquely defined equivalent control $u_{\mathrm{eq}}(x) = (CB)^{-1}CAx$ inducing sliding on S, $\tau(x) = [I - B(CB)^{-1}C]Ax$, and uniqueness of the sliding solution is obviously guaranteed. Note, however, that the existence and uniqueness of an equivalent control are not necessary for approximability, as we will show in the next section. Moreover, here, in contrast to results in [17], condition (10) depends just on the geometry of the control problem and is verifiable a priori, once the velocity field and the sliding manifold are known. Therefore the precise knowledge of the solution set $W(x_0)$ is not required, while in general the uniqueness of the sliding motion, needed, for example, in Corollary 4.1 of [17], has to be tested on the set of trajectories that can be issued from a point of the sliding manifold applying any feedback control and using the Filippov solution definition. This difference is the outcome of Theorem 2.1, which gives a characterization of $W(x_0)$ in terms of the solution set of a regular differential inclusion.

3.2. A class of well-posed sliding mode control systems. In this part of the paper we present a family of control systems satisfying the approximability property, in the spirit of Theorem 3.4 and Corollary 3.5. Also, we show that convexity of $f(x,U)$ is not required to belong to this class, in contrast to previous results, e.g., Corollary 4.1 in [17].

Let us consider the control system

$$\begin{cases} \dot{x}_1 = f_1(x), \\ \dot{x}_2 = f_2(x,u) \end{cases} \tag{12}$$

where $x = (x_1, x_2) \in \mathbb{R}^N$, $x_1 \in \mathbb{R}^{N-M}$, $x_2 \in \mathbb{R}^M$ with $M < N$, $u \in U \subset \mathbb{R}^M$, U compact, $f_1 : \mathbb{R}^N \to \mathbb{R}^{N-M}$, and $f_2 : \mathbb{R}^N \times U \to \mathbb{R}^M$. Then clearly

$$F(x) := \operatorname{co} f(x,U) = \{f_1(x)\} \times \operatorname{co} f_2(x,U)$$

for $f : \mathbb{R}^N \times U \to \mathbb{R}^N$, $f(x,u) = (f_1(x), f_2(x,u))$. Now letting $h : \mathbb{R}^{N-M} \to \mathbb{R}^M$ be a C^1 function, we define

$$s : \mathbb{R}^N \to \mathbb{R}^M, \qquad s(x) = x_2 - h(x_1) \tag{13}$$

and let the sliding surface S be the set $\{(x_1, x_2) \in \mathbb{R}^N : x_2 = h(x_1)\}$. Then we have $Js(x) = [-Jh(x_1), I_M]$, where we have used the symbol Jg for the Jacobian matrix of g and I_M is the identity matrix of dimension M. Then obviously Js is of maximal rank at any point and $T_S(x)$, the Bouligand tangent cone to S at x, coincides with the usual tangent space; i.e.,

$$T_S(x) = \{z \in \mathbb{R}^N : Js(x)z = 0\} = \ker Js(x) \quad \forall x \in S.$$

Therefore

$$\begin{aligned} F(x) \cap T_S(x) &= \{(f_1(x), v) : v \in \operatorname{co} f_2(x,U),\ Js(x)\,(f_1(x), v)^T = 0\} \\ &= \{(f_1(x), v) : v \in \operatorname{co} f_2(x,U),\ v - Jh(x_1)f_1(x) = 0\} \end{aligned}$$

for any $x = (x_1, x_2) \in S$, so that

$$F(x) \cap T_S(x) = \begin{cases} \varnothing & \text{if } Jh(x_1)f_1(x) \notin \operatorname{co} f_2(x,U), \\ \{(f_1(x), Jh(x_1)f_1(x))\} & \text{otherwise.} \end{cases}$$

Thus from Theorem 3.4 a sliding mode on S exists if and only if

$$Jh(x_1)f_1(x) \in \operatorname{co} f_2(x,U) \tag{14}$$

at any point $x = (x_1, h(x_1))$ with $x_1 \in \mathbb{R}^{N-M}$. Moreover, if this condition is satisfied, sliding modes for (12)–(13) are given by the solutions of

$$\begin{cases} \dot{x}_1 = f_1(x), & x_1(0) = \overline{x}_1, \\ \dot{x}_2 = Jh(x_1)f_1(x), & x_2(0) = h(\overline{x}_1) \end{cases}$$

or, equivalently, setting $g(x_1) = f_1(x_1, h(x_1))$,

$$\begin{cases} \dot{x}_1 = g(x_1), & x_1(0) = \overline{x}_1, \\ x_2(t) = h(x_1(t)). \end{cases}$$

Since h is C^1, if f_1 is Lipschitz continuous, the above differential equation admits a unique solution, and by Corollary 3.5, sliding mode control system (12)–(13) enjoys the approximability property (A).

In Corollary 4.1 in [17] it is proved that if $f(x,U)$ is convex and the sliding mode is unique, the approximability property is satisfied. We show here by an example that for the above class of systems convexity of the velocity field is not required.

Example 3.1. Let $N = 3$, $M = 2$, $x = (x_1, x_2, x_3)$, $u = (u_1, u_2)$ with $|u_i| \le 1$, $f_1 : \mathbb{R}^3 \to \mathbb{R}$, and

$$\begin{cases} \dot{x}_1 = f_1(x), \\ \dot{x}_2 = u_1, \\ \dot{x}_3 = u_1 u_2, \end{cases} \qquad f(x,u) = (f_1(x), u_1, u_1 u_2).$$

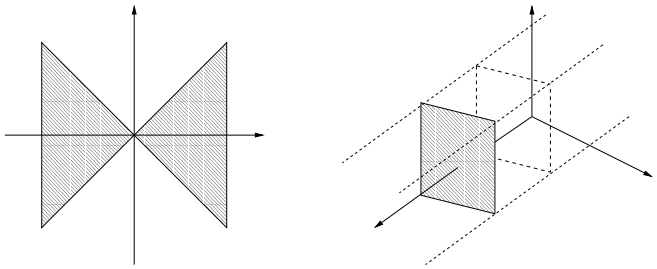


FIG. 1. *In* (A) *we depicted the set* $\{(u_1, u_1u_2) : |u_i| \le 1\}$*, while the set* $\operatorname{co} f(x, U)$ *for fixed* x *is shown in* (B)*. Note that this set always belongs to the intersection between the dashed parallelepiped and a plane orthogonal to the first axis, while for our choice of* S *the set* $T_S(x)$ *will always be a line passing through the origin.*

Then

$$f(x,U) = \left\{ f_1(x)\} \times \{(u_1, u_1u_2) : |u_i| \le 1\} = \{f_1(x)\} \times \bigcup_{|\alpha|\le 1} \{(t, \alpha t) : |t| \le 1 \right\}$$

(see Figure 1). Then $f(x,U)$ is not convex and $\operatorname{co} f(x,U) = \{f_1(x)\} \times [-1,1]^2$.

In this example we define $s : \mathbb{R}^3 \to \mathbb{R}^2$, $s(x) = (x_2, x_3) - h(x_1)$ with a regular $h : \mathbb{R} \to \mathbb{R}^2$ and $S = \{(x_1, h(x_1)) : x_1 \in \mathbb{R}\}$. Writing $h(x_1) = (h_1(x_1), h_2(x_1))$, we obtain $Jh(x_1) = (h_1'(x_1), h_2'(x_1))^T$. Letting $g_1(x_1) = f_1(x_1, h_1(x_1), h_2(x_1))$, the necessary and sufficient condition (14) for the existence of a sliding motion on S is $Jh(x_1)g_1(x_1) \in [-1,1]^2$, i.e.,

$$|h_i'(x_1)g_1(x_1)| \le 1, \qquad i = 1, 2.$$

For example, if $h_1 = h_2 = 0$, so that S is the first coordinate axis, the condition is satisfied, and the only admissible sliding motion on S is the solution of a Cauchy problem

$$\begin{cases} \dot{x}_1 = f_1(x_1, 0, 0), \\ x_1(0) = \overline{x}_1, \end{cases} \qquad x_2(t) = x_3(t) = 0.$$

Note that in this case, for any starting point on S, any choice of a control law $(0, u_2)$ with $|u_2| \le 1$ is able to maintain the motion on S; i.e., the equivalent is not uniquely defined.

4. Sliding modes, nonapproximability, and regularization. While in the first part of the paper we discussed well-posedness and found classes of sliding mode control systems satisfying the approximability property, here we focus our attention on systems which do not enjoy this property for a given surface S. In section 4.1 we discuss in detail an example, due to Izosimov[1] and studied in [16]. This system was considered by Utkin in order to show the existence of systems for which the sliding motion along the sliding manifold is not uniquely defined. It turns out in [16] that the motion on the manifold depends on the implemented sequence of real states and

[1]The same example has been dealt with in a different context in [6].

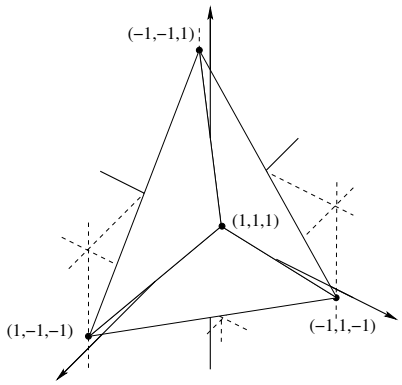


FIG. 2. $\operatorname{co} f(U)$.

that it is possible to find different ideal sliding motions passing to the limit. Here we study this example under different perspectives and we verify that in any case the system does not satisfy property (A).

A natural idea to find a regularization would be to find a sequence of manifolds S_n, converging to the starting one, for which the problem satisfies the approximability property. However, by Theorem 3.4 we show that in general this is not possible, and the aim of section 4.2 is to find an appropriate regularization of a sliding mode control system (1) which does not satisfy the approximability property. The main result is Theorem 4.1, which deals with the regularization of the integral functional corresponding to the system, as explained in section 3.1. Using this theorem, we are able to select special sequences of trajectories converging to an a priori chosen sliding state.

Finally, we make some remarks on the theorem and propose some possible applications.

4.1. A model example. We consider the following sliding mode control system (see [16], p. 35):

$$\begin{cases} \dot{x}_1 = u_1, \\ \dot{x}_2 = u_2, \\ \dot{x}_3 = u_1 u_2, \end{cases} \qquad s(t, x_1, x_2, x_3) = (x_1, x_2), \quad S = \{(0, 0, x_3) : x_3 \in \mathbb{R}\} \tag{15}$$

with $x = (x_1, x_2, x_3) \in \mathbb{R}^3$, $u = (u_1, u_2) \in U := [-1, 1]^2$, $t \in [0, 1]$, $f(t, x, u) = f(u) = (u_1, u_2, u_1 u_2)$, and initial point $x_0 \in S$.

The system defined by (15) fulfills the hypotheses stated in section 2.1. If we take relaxed controls as admissible controls, we have to consider all solutions of the differential inclusion $\dot{x}(t) \in \operatorname{co} f(U)$ and (see Figure 2)

$$\operatorname{co} f(U) = \operatorname{co}\{(1, 1, 1), (1, -1, -1), (-1, -1, 1), (-1, 1, -1)\}. \tag{16}$$

Note that in this example $\operatorname{co} f(U) = F_{\hat{u}}(x)$ for any $x \in S$ if the control $\hat{u}$ is given by

$$\hat{u}_1(t, x) = \operatorname{sign}(x_1), \qquad \hat{u}_2(t, x) = \operatorname{sign}(x_2)$$

and $F_{\hat{u}}(x)$ is the corresponding Filippov multifunction, as defined in (5). Consider now a sliding trajectory $x(t) = (0, 0, x_3(t))$. Then by Theorem 3.4, $x(\cdot)$ satisfies

$$\dot{x}(t) \in \operatorname{co} f(U) \cap T_S(x(t)), \quad \text{for a.e. } t,$$

where $T_S(x) = S$ in this case for every $x \in S$. We therefore get that the set of sliding states with initial point x_0 is

$$\{(0, 0, x_3(t)) \in AC(0, T; \mathbb{R}^3) : |\dot{x}_3(t)| \leq 1\}. \tag{17}$$

Therefore in this case the sliding trajectory is not unique, and consequently the system is not approximable. We remark that these sliding motions also were computed in [16], using a different argument.

If we take into account only Carathéodory solutions of system (15), the unique sliding state is the trajectory

$$x(t) = x_0$$

for every $t \in [0, 1]$, corresponding to the control $(u_1, u_2) = (0, 0)$.

Let us show, however, that this does not imply approximability. Choose for instance $x_0 = (0, 0, 0)$ and consider the sequence of controls

$$u_1^n(t) = \begin{cases} 1 & \text{if } t \in I_k^n, \\ -1 & \text{if } t \in D_k^n, \end{cases} \qquad u_2^n = u_1^n$$

and the corresponding trajectories solving system (15), starting from x_0,

$$x_1^n(t) = \begin{cases} t - \frac{2k}{2^n} & \text{if } t \in I_k^n, \\ -t + \frac{2k}{2^n} & \text{if } t \in D_k^n, \end{cases} \qquad x_2^n = x_1^n, \qquad x_3^n(t) = t,$$

where $I_k^n = [2k/2^n, (2k+1)/2^n]$ for $k \in \{0, \ldots, 2^{n-1}-1\}$ and $D_k^n = [(2k-1)/2^n, 2k/2^n]$ for $k \in \{1, \ldots, 2^{n-1}\}$. These trajectories are approaching the sliding manifold, because (x_1^n, x_2^n) is uniformly convergent to $(0, 0)$, but the whole sequence is convergent to $x(t) = (0, 0, t)$, which is not the unique Carathéodory sliding state. Thus the system does not satisfy property (A). Note that, in the same way, it is possible to construct sequences of Carathéodory real sliding states approaching any of the relaxed sliding solutions we found before.

The aim of the next section is to find a method to regularize such a system. One possibility would be to find a sequence of surfaces S_n such that

(S1) the sliding mode control defined by (15) with sliding manifold S_n satisfies property (A) for each n;

(S2) $S_n \to S$ in some sense;

(S3) S_n is of the "same kind" of S.

However, in this example, it is not possible. Consider condition (S3) and suppose we look for sliding manifolds S_n such that

$$S_n = \{x \in \mathbb{R} : s_n(x) = 0\}$$

with $s_n : \mathbb{R}^3 \to \mathbb{R}^2$ a C^2 function with Jacobian matrix $Js_n(x)$ of maximal rank at any point $x \in S_n$. With these assumptions, it follows that $T_{S_n}(x) = \{y \in \mathbb{R}^3 : Js_n(x)y = 0\}$. Regardless of the choice of S_n, the cone $T_{S_n}(x)$ is a line passing through the origin for every $x \in S_n$. Recall that here $\operatorname{co} f(U)$ does not depend on x and is the polyhedral set (16), containing the origin (see Figure 2). Thus $T_{S_n}(x) \cap \operatorname{co} f(U)$ is a segment of positive length. According to Theorem 3.4, every $x \in AC(0, T)$ solving

$$\begin{cases} \dot{x}(t) \in T_{S_n}(x(t)) \cap \operatorname{co} f(U) & \text{for a.e. } t, \\ x(0) \in S_n \end{cases}$$

is a trajectory sliding on S_n. Since the right-hand side admits more than one continuous selection, by Corollary 3.5, property (A) is not satisfied, regardless of the choice of S_n. Therefore requirements (S1) and (S3) are not compatible, no matter what convergence is chosen in (S2).

4.2. Regularization. As in the model problem, if a sliding mode exists, nonapproximability is the consequence of the lack of uniqueness of the sliding state (Corollary 3.2). Our aim here is to find a method to select sequences of real states converging to a prescribed sliding motion. Relying on the equivalence result in Theorem 3.1, what we need is to devise a way to appropriately select minimizing sequences of the ill-posed functional I,

$$I(x) = \int_0^T |s(t, x(t))|\, dt, \tag{18}$$

on the set $W(x_0)$. This can be done by using regularization processes such as Tikhonov's (see [11]). A sequence of functionals I_n is defined by adding a small uniformly convex term to I so that $(W(x_0), I_n)$ is well-posed for all n. Uniform convexity then assures that minimizing sequences for I can be generated through I_n and convergence to a fixed minimizer of I is guaranteed. In our setting, the compactness property of $W(x_0)$ allows us to weaken the required regularity of the perturbations and to obtain the following new regularization result.

THEOREM 4.1. *Assume that the set* $P := \operatorname{argmin}(W(x_0), I)$ *is nonempty and let* $J, J_n : W(x_0) \to \mathbb{R}$ *be lower semicontinuous functionals such that*

(i) J *has a unique minimizer* $\overline{y}$ *on the set* P*; i.e.,* $\operatorname{argmin}(P, J) = \{\overline{y}\}$ *and* $J(\overline{y}) = 0$*;*

(ii) $J(y) \le \liminf J_n(y_n)$ *for every* y *and any sequence* y_n *such that* $y_n \to y$ *uniformly;*

(iii) *there exists* $\overline{z} \in P$ *such that* $\limsup J_n(\overline{z}) = 0$.

Consider sequences $a_n > 0$, $\varepsilon_n \ge 0$ *such that* $a_n \to 0$, $\varepsilon_n \to 0$, *and* $(\varepsilon_n/a_n) \to 0$, *and define* $I_n = I + a_n J_n$. *Then for every sequence* $y_n \in W(x_0)$ *such that* $I_n(y_n) \le \varepsilon_n + \inf I_n$, *it follows that*

(a) y_n *is a minimizing sequence for* I*;*

(b) $y_n \to \overline{y}$.

Proof. For any $n \in \mathbb{N}$, let $y_n \in W(x_0)$ be such that $I_n(y_n) \le \epsilon_n + \inf I_n$. By the compactness property proved in Lemma 2.2, up to subsequences, y_n is convergent to a certain $y^* \in W(x_0)$.

From the definition it follows that

$$I(y_n) = I_n(y_n) - a_n J_n(y_n) \le \epsilon_n + \inf I_n - a_n J_n(y_n); \tag{19}$$

by assumption (ii), $\liminf J_n(y_n) \ge J(y^*)$ and it can be easily proved that $\inf I_n \to \inf I = 0$ (see Remark 4.1). Thus, passing to the upper limit in (19), we get

$$\limsup I(y_n) \le 0; \tag{20}$$

therefore y_n is minimizing for I. Moreover, since I is continuous, we obtain $I(y^*) = 0$, i.e., $y^* \in P$; therefore to show that $y^* = \overline{y}$, it is enough to prove that y^* is a minimizer of J on P. Since $I \ge 0$ and $a_n > 0$, from (19) we get

$$J_n(y_n) \le J_n(y_n) + \frac{I(y_n)}{a_n} \le \frac{\varepsilon_n}{a_n} + \frac{I_n(\overline{z})}{a_n} = \frac{\varepsilon_n}{a_n} + J_n(\overline{z}). \tag{21}$$

Thus, by assumptions (ii) and (iii), since $(\varepsilon_n/a_n) \to 0$, we have

$$0 \le J(y^*) \le \liminf J_n(y_n) \le \limsup J_n(y_n) \le 0. \tag{22}$$

Putting together (20) and (22), from assumption (i) it follows that $y^* = \overline{y}$. ∎

Remark 4.1. Assumptions (i), (ii), and (iii) imply variational convergence of the sequence J_n to J. Therefore from [11, Theorem 5, p. 122] it follows that $\lim_n \inf I_n = \inf I$.

Example 4.1. Consider the nonapproximable sliding mode control system defined in section 4.1 with $x_0 = (0, 0, 0)$. We show how to apply Theorem 4.1 in order to select suitable minimizing sequences converging to a fixed sliding motion.

Since by (17) we have $P = \{(0, 0, x_3(t)) : |\dot{x}_3(t)| \le 1,\ x_3(0) = 0\}$, a feasible choice for J is

$$J(y) = \int_0^1 |y_3(t)|\, dt.$$

This functional satisfies the hypotheses of Theorem 4.1; in fact it is lower semicontinuous on $W(x_0)$ with respect to the uniform convergence, since it is lower semicontinuous on the whole space $AC(0, T; \mathbb{R}^3)$ and it has the unique constant minimizer $x(t) = x_0$ on the set of sliding states P. Then we can either take $J_n = J$ for all n or define

$$J_n(y) = \int_0^1 g_n(y_3(t))\, dt \quad \text{with} \quad g_n(y_3) = \begin{cases} |y_3| & \text{if } |y_3| \le n, \\ 0 & \text{otherwise.} \end{cases}$$

In both cases, choosing $\overline{z} = x$, condition (iii) is satisfied. Note also that J is not uniformly convex, so that J is not an admissible perturbation of the Tikhonov type.

Supposing $J_n = J$, which kind of controls are then required in order to obtain convergent asymptotically minimizing trajectories? Suppose that u_n is an admissible control law such that the corresponding trajectory y_n satisfies $I_n(y_n) \le \varepsilon_n$. Since $\epsilon_n/a_n \to 0$, from (15), it follows that

$$\frac{u_1^n}{a_n} \to 0; \quad \frac{u_2^n}{a_n} \to 0; \quad u_1^n u_2^n \to 0 \text{ weakly in } L^2(0, T). \tag{23}$$

Therefore $\|u_2^n\| \le M a_n$ for some $M > 0$; i.e., $u_2^n \to 0$ in the L^2 norm. The same holds for u_1^n. This is therefore a necessary condition that "regularizing" controls must satisfy.

Conversely, let u_n fulfill condition (23) and let y_n be a corresponding solution of (15) with starting point "near" S. Then, by standard arguments we get that $y_n \to 0$ pointwise. By the compactness property proved in Lemma 2.2, we get that actually $y_n \to 0$ uniformly.

Consider an arbitrary sliding mode control system satisfying the hypotheses stated in section 2.1. It is always possible to select sequences of real states converging to a chosen sliding state $\overline{y}$ by choosing the distance functional, namely,

$$J(y) = J_n(y) = \int_0^T |y(t) - \overline{y}(t)|\, dt.$$

There is at least another possible choice when an equivalent control u_{eq} exists.

COROLLARY 4.2. *Let the hypotheses of the general setting of section* 2.1 *be satisfied. Suppose that there exists an equivalent control* $u_{\mathrm{eq}} \in \mathcal{U}$*, i.e., an admissible feedback satisfying*

$$\frac{\partial s}{\partial t} + J_x s(t,x) f(t,x,u_{\mathrm{eq}}(t,x)) = 0 \qquad \forall x \in S(t), \quad \text{for a.e. } t \in [0,T].$$

Suppose, moreover, that the corresponding sliding trajectory y_{eq} *is unique. Let*

$$J^{\mathrm{eq}}(y) = \int_0^T \left| y(t) - x_0 - \int_0^t f(s, y(s), u_{\mathrm{eq}}(s, y(s)))\, ds \right| dt.$$

Then the sequence $J_n = J^{\mathrm{eq}}$ *for any* n *fulfills the assumptions of Theorem* 4.1.

When J_n are integral functionals satisfying suitable hypotheses, it is possible, in some sense, to read Theorem 4.1 as a result in the direction outlined at the end of the previous section. In fact if $J_n(y) = \int_0^T g_n(t, y(t))\, dt$, then

$$I_n(y) = \int_0^T \left[|s(t, y(t))| + a_n g_n(t, y(t)) \right] dt.$$

If the integrand is positive, there exists $s_n : [0,T] \times \mathbb{R}^N \to \mathbb{R}^M$ such that $|s_n(t,y)| = |s(t,y)| + a_n g_n(t,y)$, and we can think of the sets of zeros of s_n as an approximating sequence for the sliding manifold. This is not in contrast with the outcomes of section 4.1. In fact, there we showed that the sliding mode control system is not approximable whenever the sliding manifold is a fixed regular surface, while here we see that this problem can be overcome by allowing time-dependent manifolds.

REFERENCES

[1] Z. ARTSTEIN, *Rapid oscillations, chattering systems, and relaxed controls*, SIAM J. Control Optim., 27 (1989), pp. 940–948.

[2] Z. ARTSTEIN, *Chattering variational limits of control systems*, Forum Math., 5 (1993), pp. 369–403.

[3] J.-P. AUBIN, *Viability theory*, Systems Control Found. Appl., Birkhäuser Boston, Boston, MA, 1991.

[4] J.-P. AUBIN AND A. CELLINA, *Differential Inclusions*, Grundlehren Math. Wiss. 264, Springer-Verlag, Berlin, 1984.

[5] J.-P. AUBIN AND H. FRANKOWSKA, *Set-Valued Analysis*, Systems Control Found. Appl. 2, Birkhäuser Boston, Boston, MA, 1990.

[6] G. BARTOLINI, E. PUNTA, AND T. ZOLEZZI, *Second order approximability for sliding mode control systems*, in Proceedings of the 8th International Workshop on Variable Structure Systems, Vilanova i la Geltru', Spain, 2004.

[7] G. BARTOLINI, E. PUNTA, AND T. ZOLEZZI, *First and second order sliding mode regularization techniques: The approximability property*, in Proceedings of the 16th IFAC World Congress, Prague, Czech Republic, 2005.

[8] G. BARTOLINI AND T. ZOLEZZI, *Control of nonlinear variable structure systems*, J. Math. Anal. Appl., 118 (1986), pp. 42–62.

[9] G. BARTOLINI AND T. ZOLEZZI, *Behavior of variable-structure control systems near the sliding manifold*, Systems Control Lett., 21 (1993), pp. 43–48.

[10] K. DEIMLING, *Multivalued Differential Equations*, de Gruyter Ser. Nonlinear Anal. Appl. 1, Walter de Gruyter, Berlin, 1992.

[11] A. L. DONTCHEV AND T. ZOLEZZI, *Well-Posed Optimization Problems*, Lecture Notes in Math. 1543, Springer-Verlag, Berlin, 1993.

[12] C. EDWARDS AND S. K. SPURGEON, *Sliding Mode Control: Theory and Applications*, The Taylor and Francis Systems and Control Book Series, Taylor and Francis, London, 1998.

[13] H. O. FATTORINI, *Infinite-Dimensional Optimization and Control Theory*, Encyclopedia Math. Appl. 62, Cambridge University Press, Cambridge, UK, 1999.

[14] P. Nistri and M. Quincampoix, *On open-loop and feedback attainability of a closed set for nonlinear control systems*, J. Math. Anal. Appl., 270 (2002), pp. 474–487.
[15] R. T. Rockafellar, *Integral functionals, normal integrands, and measurable selections*, in Nonlinear Operators and the Calculus of Variations (Summer School, Univ. Libre Bruxelles, Brussels, 1975), Lecture Notes in Math. 543, Springer-Verlag, Berlin, 1976, pp. 157–207.
[16] V. I. Utkin, *Sliding Modes in Control and Optimization*, Comm. Control Engrg. Ser., Springer-Verlag, Berlin, 1992.
[17] T. Zolezzi, *Well-posedness and sliding mode control*, ESAIM Control Optim. Calc. Var., 11 (2005), pp. 219–228.

SIAM J. OPTIM.
Vol. 18, No. 3, pp. 895–906

DISCRETE APPROXIMATIONS AND FIXED SET ITERATIONS IN BANACH SPACES*

TZANKO DONCHEV†, ELZA FARKHI‡, AND SIMEON REICH§

Abstract. We study autonomous differential inclusions with right-hand sides satisfying a one-sided Lipschitz (OSL) condition in Banach spaces with uniformly convex duals. We first show that the solution set is closed and obtain estimates for Euler-type discrete approximations. We then use these results to derive an analogue of the exponential formula for the reachable set, as well as results regarding the existence and approximation of a strongly invariant attractor in the case of a negative OSL constant. As a by-product, conditions for controllability of the reverse-time system are obtained.

Key words. Banach space, differential inclusion, duality mapping, Euler approximations, exponential formula, fixed set iterations, one-sided Lipschitz condition, upper hemicontinuous

AMS subject classifications. 34A60, 34A45, 49J24

DOI. 10.1137/060659326

1. Introduction and preliminaries.

We study the initial value problem

$$\dot{x}(t) \in F(x(t)),\ x(0) = x_0,\ t \in I, \tag{1.1}$$

in a Banach space E with a uniformly convex dual, where I is either the bounded interval $[0, T]$ or $[0, \infty)$. A solution is any absolutely continuous function $x(\cdot)$ satisfying (1.1) for a.a. t.

Here $F : E \rightrightarrows E$ is a multifunction with nonempty, convex and weakly compact values. We assume that $F(\cdot)$ is one-sided Lipschitz (OSL), possibly discontinuous, and, in the case of an infinite time interval, that the OSL constant is negative.

The OSL condition for multifunctions with values in $\mathbf{R}^n$ is studied in [7, 8, 9, 10, 17] and in infinite-dimensional spaces in [5, 6]. Under this condition, which is considerably weaker than the classical Lipschitz one, we establish in the present paper new results on the existence of solutions and stability of the solution set with respect to perturbations of the right-hand side and the initial conditions, as well as the strict contractivity of the reachable set mapping when the OSL constant is negative.

In the case of an infinite time domain, we obtain the existence of a strongly invariant attractor set which is the limit of Euler-type set iterations. These results extend those obtained in [9] for the finite-dimensional case. Our qualitative results are obtained using Euler-type discrete approximations of (1.1) defined in [11].

The method of discrete approximations is frequently used to obtain existence of solutions of differential equations and inclusions (see, e.g., [3, 4, 12, 16]) as well as

*Received by the editors May 8, 2006; accepted for publication (in revised form) March 17, 2007; published electronically October 4, 2007.
http://www.siam.org/journals/siopt/18-3/65932.html

†Department of Mathematics, University of Architecture and Civil Engineering, 1 "Hr. Smirnenski" St., 1046 Sofia, Bulgaria (tzankodd@gmail.com).

‡School of Mathematical Sciences, Sackler Faculty of Exact Sciences, Tel Aviv University, 69978 Tel Aviv, Israel (elza@post.tau.ac.il). This author was partially supported by the Minkowski Center for Geometry at Tel Aviv University.

§Department of Mathematics, The Technion—Israel Institute of Technology, 32000 Haifa, Israel (sreich@techunix.technion.ac.il). This author was partially supported by the Fund for the Promotion of Research at the Technion and by the Technion VPR Fund.

necessary optimality conditions [18]. We also mention [13], where discrete approximations of singularly perturbed systems are studied. Applications of infinite-dimensional differential equations and inclusions are presented in a number of papers and books; see, e.g., [15, 19]. In particular, suitable modifications of the partial differential inclusion described in [14, section I.7] and of the partial differential equation (4.1) in [19] can serve as examples of applications of our results.

We also remark in passing that descent methods for minimizing a (convex) function f,

$$x_{k+1} = x_k + h_k f_k,\ f_k \in F(x_k) = -\partial f(x_k),$$

where ∂f is the subdifferential of f, may be regarded as Euler-type discrete approximations of (1.1) . Under certain conditions on f, such iterations approach an invariant set for (1.1) containing the minimizers of f.

Our paper is organized as follows. In the next section a semidiscrete approximation of (1.1) is defined and error estimates for the solution and reachable sets, as well as the nonemptiness and closedness of the solution set of (1.1), are derived. In the third section these results are applied to obtain the main results of the paper on the Euler approximation of (1.1) for finite and infinite time intervals. In the last section we present conditions for controllability of the reverse-time system.

We now recall some definitions and notation and refer the reader to [3, 14, 15] for all concepts used in this paper, but not explicitly discussed here. Define the normalized duality mapping by $J(x) = \{l \in E^* : \langle l, x\rangle = |x|^2 = |l|^2\}$, $x \in E$. Since E^* is uniformly convex, $J(\cdot)$ is single-valued and uniformly continuous on bounded sets (see [3, 15]). For $M > 0$ we denote by $\omega(\delta, M) := \sup\{|J(x) - J(y)| :\ |x - y| \le \delta,\ |x| \le M,\ |y| \le M\}$ its modulus of continuity on the ball $\{x \in E : |x| \le M\}$.

The support function of the bounded set A is denoted by $\sigma(y, A) := \sup_{x\in A}\langle x, y\rangle$ and the Hausdorff distance between the (closed and bounded) sets A and C by $D_H(A, C) = \max\{ex(A, C), ex(C, A)\}$, where $ex(A, C) = \sup_{a\in A}\inf_{b\in C}|a - b|$. The multifunction $R(\cdot)$ is said to be upper hemicontinuous (UHC) when for every $l \in E^*$ the support function $\sigma(l, R(\cdot))$ is USC as a real-valued function. Also, $F(\cdot)$ is said to be lower semicontinuous (LSC) at y when for every $f \in F(y)$ and every $y_i \to y$ there exists $f_i \in F(y_i)$ with $f_i \to f$. The (multi)function $R : I \rightrightarrows E$ is said to be strongly measurable if for every bounded interval $[a, b]$ and every $\varepsilon > 0$, there exists a simple function $R_i(\cdot)$ such that $D_H(R_i(t), R(t)) < \varepsilon$ for a.a. $t \in [a, b]$. Equivalently, $R(\cdot)$ is strongly measurable if for every bounded interval $[a, b]$ and every $\varepsilon > 0$, there exists a compact I_ε with $\mathrm{meas}(I_\varepsilon) > b - a - \varepsilon$ such that $R(\cdot)$ is continuous on I_ε (with respect to the Hausdorff distance).

By $diam(C) = \sup\{|a - b| :\ a, b \in C\}$ we denote the diameter of the set C, $|C| = D_H(C, \{0\})$ is the "norm" of a bounded set, and by $\mathbb{B}$ we denote the open unit ball in E.

We let $\mathrm{Graph}_A F := \{(x, y) :\ x \in A,\ y \in F(x)\}$ be the graph of F on a set A.

DEFINITION 1.1. *The multifunction $F :\ E \rightrightarrows E$ is said to be OSL with a constant L (not necessarily positive) when*

$$\sigma(J(x - y), F(x)) - \sigma(J(x - y), F(y)) \le L|x - y|^2$$

for every $x, y \in E$.

In this paper we assume that $L < 0$ when $T = \infty$.

Given a partition $\Delta = \{0 = t_0 < t_1 < \cdots < t_n = T\}$, we set $h_\Delta = \max_{1\le i\le n}(t_i - t_{i-1})$ (when $T = \infty$ we assume that $n = \infty$ and replace max by sup). The semidiscrete

differential inclusion corresponding to (1.1) is

$$\dot{x}(t) \in F(x_i),\ x(0) = x_0,\ x_i = x(t_i),\ t \in [t_i, t_{i+1}],\ i = 0, 1, \ldots, n-1. \tag{1.2}$$

Let $A(t, x_0)$ be the reachable set of (1.1) at the time t. The reachable set of (1.2) at t_k will be denoted by $A^{\Delta}(k, x_0)$ or by $A^{\Delta}(k)$ when x_0 is fixed. We denote by R_2^{Δ} the solution set of (1.2) and by R_1 the solution set of (1.1).

In the next section we study the connections between the solution set of (1.1) and the solution set of (1.2) (the set of discrete trajectories of (1.1)).

We will use the following simple inequality:

$$\big||a|^2 - |b|^2\big| \le |a-b|\{|a| + |b|\}. \tag{1.3}$$

2. Approximation of the solution set. In this section we estimate the Hausdorff distance between the solution sets of (1.1) and (1.2) and their reachable sets. The more restrictive case of a compact-valued (and USC) F is studied in [4].

In particular, we present more general versions of some results of [6], when E is not a Hilbert but a Banach space with a uniformly convex dual. We also extend to infinite-dimensional spaces the main results of [9] and [10] .

We will make the following assumption.

A. *$F(\cdot)$ is bounded on bounded sets and UHC.*

LEMMA 2.1. *Let F be an OSL multifunction which satisfies assumption* A. *Then there exist constants M and N such that $|x(t)| \le M$ and $|F(x(t) + \mathbb{B})| \le N$ for a.e. $t \in I$ and every solution $x(\cdot)$ of*

$$\dot{x}(t) \in \overline{co}\ F(x(t) + \mathbb{B}) + \mathbb{B},\ x(0) \in x_0 + \mathbb{B}. \tag{2.1}$$

Proof. This result is proved when F is also compact-valued in [4]. The proof in our case is similar but is included here for the convenience of the reader. Let $R(x) := \overline{co}\ F(x + \mathbb{B}) + \mathbb{B}$. Note that $R(\cdot)$ is also OSL with constant L, and bounded on bounded sets. Indeed, $\sigma(l, F(x) + \mathbb{B}) = \sigma(l, F(x)) + \sigma(l, \mathbb{B})$ for every $l \in E^*$. Furthermore, for every $x, y \in E$ and every $\varepsilon > 0$, there exists $l_\varepsilon \in \mathbb{B}$ such that $\sigma(J(x-y), F(x+\mathbb{B})) \le \sigma(J(x-y), F(x+l_\varepsilon)) + \varepsilon$. Thus $\sigma(J(x-y), F(x+\mathbb{B})+\mathbb{B}) - \sigma(J(x-y), F(y+\mathbb{B})+\mathbb{B}) \le \sigma(J(x-y), F(x+l_\varepsilon)) + \varepsilon - \sigma(J(x-y), F(y+l_\varepsilon)) \le L|x-y|^2 + \varepsilon$. Hence the mapping $x \to \overline{co}\ F(x+\mathbb{B}) + \mathbb{B}$ is also OSL with constant L.

If $x(\cdot)$ is a solution of (2.1), then

$$\begin{aligned}
&\langle J(x(t)), \dot{x}(t) - 0\rangle \le \sigma(J(x(t)), R(x(t))) - \sigma(J(x(t)), R(0)) \\
&+\sigma(J(x(t)), R(0)) \le L|x(t)|^2 + |R(0)| \cdot |x(t)|,\ \text{i.e.,} \\
&\frac{d}{dt}|x(t)|^2 \le 2L|x(t)|^2 + 2|R(0)| \cdot |x(t)|.
\end{aligned}$$

Since $|x(\cdot)|$ is absolutely continuous, it is a.e. differentiable and, moreover, $\frac{d}{dt}|x(t)|^2 = 2|x(t)|\frac{d}{dt}|x(t)|$. Let $\mathcal{J} := \{t :\ |x(t)| = 0\}$. If $s \in \mathcal{J}$ is a point of density where $|x(\cdot)|$ is differentiable, then $\frac{d}{ds}|x(s)| = 0$. Thus $\frac{d}{dt}|x(t)| \le L|x(t)| + |R(0)|$ for a.a. $t \in I$. But $|R(0)| \le M'$. Hence there exists a number M such that $M \ge |x(t)|$ for all $t \in I$ and for all solutions $x(\cdot)$ of (2.1). Since R is bounded on bounded sets, we can also find $N \ge |R(M\mathbb{B})|$. □

It is easy to see that $|x(t)| \le e^{Lt}(|x_0| + \int_0^t e^{-Ls}|R(0)|ds)$. Suppose $L \ne 0$. Then $|x(t)| \le e^{Lt}|x_0| + \frac{e^{Lt}-1}{L}|R(0)|$. Hence, in case $L < 0$, the constants M and N do not depend on the time $t > 0$.

Once M is determined, we let $\omega_J(\delta) = \omega_J(\delta, M)$.

We will use the following properties of $\omega_J(\cdot)$:

(a) The modulus $\omega_J(\cdot)$ is increasing.

(b) If $K \geq 1$, then $\omega_J(K\delta) \leq K\omega_J(\delta)$.

Indeed, given $\varepsilon > 0$, let $|J(x) - J(y)| \geq \omega_J(K\delta) - \varepsilon$. Therefore,

$$\left|J\left(\frac{x}{K}\right) - J\left(\frac{y}{K}\right)\right| = \frac{1}{K}|J(x) - J(y)|.$$

However, one has $\left|J\left(\frac{x}{K}\right) - J\left(\frac{y}{K}\right)\right| \leq \omega_J\left(\delta, \frac{M}{K}\right) \leq \omega_J(\delta)$, i.e., $\omega_J(K\delta) - \varepsilon \leq K\omega_J(\delta)$. Since ε is arbitrary, the result follows.

From now on we assume that $F(\cdot)$ has convex values.

THEOREM 2.2. *If $F(\cdot)$ is OSL and satisfies assumption* A*, then R_1 is nonempty and closed, and there exists a constant $C = C(T)$ such that*

$$D_H(R_2^{\Delta}, R_1) \leq C\sqrt{h_\Delta + \omega_J(h_\Delta)}.$$

If $L < 0$, then $C(\infty) < \infty$.

Proof. Let $\dot{y}(t) \in F(y(t) + \varepsilon\mathbb{B})$, i.e., $\dot{y}(t) \in F(y(t) + h(t))$, where $|h(t)| \leq \varepsilon$, and let $\Delta = \{t_i\}_{i=1}^n$ be a partition of $[0, T]$. We look for a solution $x(\cdot)$ of the discrete inclusion which is close to $y(\cdot)$. Let $x(0) = x_0$. Assuming that $x(\cdot)$ is known on $[0, t_i]$, we find $x(\cdot)$ on $[t_i, t_{i+1}]$. Denote $x_i = x(t_i)$. The OSL condition implies that

$$\sigma(J(y(t) - x_i + h(t)), F(y(t) + h(t))) - \sigma(J(y(t) - x_i + h(t)), F(x_i)) \leq L|y(t) - x_i + h(t)|^2.$$

It is easy to see that $\sigma(J(y(t)-x_i+h(t)), F(x_i)) \leq N\omega_J(\varepsilon)+\sigma(J(y(t)-x_i), F(x_i))$ and that $\sigma(J(y(t) - x_i + h(t)), F(y(t) + h(t))) \geq \sigma(J(y(t) - x_i), F(y(t) + h(t))) - N\omega_J(\varepsilon)$.

Let $\tilde{z}(t) = y(t) - x_i$. Since the mapping $t \to \tilde{z}(t)$ is continuous, it follows that for every $\delta > 0$, there exists a strongly measurable selection $f_i(t) \in F(x_i)$ $(x_i = x(t_i))$ on the interval $[t_i, t_{i+1}]$ such that

$$\sigma(J(\tilde{z}(t)), F(x_i)) < \langle J(\tilde{z}(t)), f_i(t)\rangle + \delta.$$

There is even such a continuous selection because for every $\mu > 0$ the set-valued map $G_\mu(t) = \{\overline{y \in F(x_i) : \langle J(\tilde{z}(t)), y\rangle > \sigma(J(\tilde{z}(t)), F(x_i)) - \mu}\}$ is LSC with nonempty, convex, and weakly compact values.

Setting $x(t) = x_i + \int_{t_i}^t f_i(\tau)\, d\tau$ and $z(t) = y(t) - x(t)$, we obtain

$$\sigma(J(\tilde{z}(t)), F(y(t) + h(t))) - \sigma(J(\tilde{z}(t)), F(x_i)) \leq 2N\omega_J(\varepsilon) + L|\tilde{z}(t) + h(t)|^2$$

and

$$L|\tilde{z}(t) + h(t)|^2 \leq L|z(t)|^2 + |L|\left||\tilde{z} + h(t)|^2 - |z(t)|^2\right|.$$

From the triangle inequality and (1.3) it follows that

$$\begin{aligned}(2.2)\qquad &\sigma(J(\tilde{z}(t)), F(y(t) + h(t))) - \sigma(J(\tilde{z}(t)), F(x_i)) \\ &\leq 2N\omega_J(\varepsilon) + L|z(t)|^2 + |L||\tilde{z}(t) + h(t) - z(t)|\left(|\tilde{z}(t)| + |z(t)| + \varepsilon\right).\end{aligned}$$

Thus we have

$$\begin{aligned}&\langle J(z(t)), \dot{z}(t)\rangle \leq \langle J(\tilde{z}(t)), \dot{z}(t)\rangle + |J(z(t)) - J(\tilde{z}(t))||\dot{z}(t)| \\ &\leq \sigma(J(\tilde{z}(t)), F(y(t) + h(t))) - \sigma(J(\tilde{z}(t)), F(x_i)) + \delta + 2N\omega_J(|x(t) - x_i|).\end{aligned}$$

Since $\langle J(z(t)), \dot{z}(t)\rangle = \frac{1}{2}\frac{d}{dt}|y(t)-x(t)|^2$, taking into account (2.2) we derive

$$\frac{d}{dt}|y(t)-x(t)|^2 \le 2L|y(t)-x(t)|^2 + 4N^2\omega_J(h_\Delta) + 4N\omega_J(\varepsilon)$$
$$+2|L||x(t)-x_i+h(t)|(4M+\varepsilon) + 2\delta.$$

Hence $|y(t)-x(t)|^2 \le r(t)$, where $r(0)=0$ and

$$\dot{r}(t) = 2Lr(t) + 2|L|\,(\varepsilon + Nh_\Delta)\,(4M+\varepsilon) + 4N\omega_J(\varepsilon) + 4N^2\omega(h_\Delta) + 2\delta.$$

Let $S(T) := \max_{t\in[0,T]} e^{2Lt}\int_0^t e^{-2L\tau}d\tau$. Clearly, $S(\infty) < \infty$ for $L < 0$. Hence there exists a constant $C(T) = S(T)C_1(M,N)$ such that

$$|y(t)-x(t)| \le C(T)\sqrt{h_\Delta + \varepsilon + \omega_J(\varepsilon + h_\Delta) + 2\delta}. \tag{2.3}$$

Clearly, if $L < 0$, then $C(\infty) < \infty$.

Since $\delta, \varepsilon > 0$ are arbitrary, we get $ex(R_1, R_2^\Delta) \le C\sqrt{h_\Delta + \omega_J(h_\Delta)}$.

Now we are to prove that the solution set of (1.1) is nonempty and $C(I,E)$ closed, and that $ex(R_2^\Delta, R_1) \le C\sqrt{h_\Delta + \omega_J(h_\Delta)}$. To this end, we construct a Cauchy sequence of semidiscrete trajectories.

Let a partition $\Delta_1 = \{t_i\}_{i=1}^K$ with step h_1 be given, and let $\Delta_2 = \{\tau^j\}_{j=1}^P$ be a refinement of Δ_1 with step $h_2 \le \frac{h_1}{4}$, satisfying $\omega_J(h_2) \le \frac{\omega_J(h_1)}{4}$. Denote by $R_2^{\Delta_1}$ and $R_2^{\Delta_2}$ the solution set of (1.2) with respect to Δ_1 and Δ_2. For a given $x(\cdot) \in R_2^{\Delta_1}$, we define $y(\cdot) \in R_2^{\Delta_2}$ as follows.

Suppose $y(\cdot)$ is already defined on $[0,\tau^j]$. Let $[\tau^j, \tau^{j+1}] \subset [t_i, t_{i+1}]$. Put $y^j = y(\tau^j)$ and $x_i = x(t_i)$. For $t \in [\tau^j, \tau^{j+1}]$, we find a strongly measurable $f(t) \in F(y^j)$, where $y^j = y(\tau^j)$, such that

$$\langle J(x_i - y^j), \dot{x}(t) - f(t)\rangle \le L|x_i - y^j|^2.$$

Indeed, one can take $f(t) = l^j$, where $\langle J(x_i-y^j), l^j\rangle = \sigma(J(x_i-y^j), F(y^j))$. Define $y(t) = y^j + \int_{\tau^j}^t f(s)\,ds$. Similarly, we obtain

$$\langle J(x(t)-y(t)), \dot{x}(t)-\dot{y}(t)\rangle \le L|x(t)-y(t)|^2$$
$$+|J(x_i-y^j) - J(x(t)-y(t))||\dot{x}(t)-\dot{y}(t)| + |L|\left||x_i-y^j|^2 - |x(t)-y(t)|^2\right|.$$

Since $|F(x(t)+\mathbb{B})| \le N$ and $|x(t)| \le M$, we have

$$|(x_i-y^j) - (x(t)-y(t))| \le N(h_1+h_2); \quad |\dot{x}(t)-\dot{y}(t)| \le 2N.$$

Hence

$$\langle J(x(t)-y(t)), \dot{x}(t)-\dot{y}(t)\rangle \le L|x(t)-y(t)|^2 + 2N\omega_J(N(h_1+h_2)) + |L|4MN(h_1+h_2).$$

By property (b) of ω_J, $\omega_J(N(h_1+h_2)) \le N\omega_J(h_1+h_2)$. Denoting $r(t) = |x(t)-y(t)|^2$, one derives

$$\dot{r}(t) \le 2Lr(t) + 4N^2\omega_J(h_1+h_2) + 8MN|L|(h_1+h_2).$$

To complete the proof, one has to consider an appropriate sequence of partitions $\{\Delta_k\}_{k=1}^\infty$ with steps $h_{k+1} \le \frac{h_k}{4} \le \frac{h_1}{4^k}$, $\omega_J(h_{k+1}) \le \frac{\omega_J(h_k)}{4}$, and the corresponding

sequence of approximate solutions $\{x_k(\cdot)\}_{k=1}^{\infty}$ such that $|x_k(t) - x_{k+1}(t)|^2 = r_k(t)$. Here

$$\dot{r}_k(t) \le 2Lr_k(t) + 4N^2\omega_J\left(\frac{5h_k}{4}\right) + 10MN|L|h_k$$

and

$$r_k(t) \le \frac{Ce^{2Lt}t}{4^k}(h_1 + \omega_J(h_1)).$$

Hence $\sum_{k=1}^{\infty}\sqrt{r_k(t)}$ converges uniformly on $[0,T]$. Consequently, $\{x_k(\cdot)\}_{k=1}^{\infty}$ is a Cauchy sequence and there exists $x(t) = \lim_{k\to\infty} x_k(t)$. Since F is UHC, it is standard to show that $x(\cdot)$ is a solution of (1.1). The same method proves that the solution set is closed.

Let $s_n := \max_{t\in[0,T]}\sqrt{r_n(t)}$. Then

$$Ex(R_2^{\Delta}, R_1) \le \sum_{k+1}^{\infty} s_n \le C(T)\sqrt{h_{\Delta} + \omega_J(h_{\Delta})}.$$

Obviously, $C(\infty) < \infty$ when $L < 0$. □

The following corollary is a lemma of Filippov–Pliss type.

COROLLARY 2.3. *If $x(\cdot)$ is a solution of*

$$\dot{x}(t) \in F(x(t) + \varepsilon\mathbb{B}),\ x(0) = x_0,$$

then there exists a constant $C = C(T)$ such that $\operatorname{dist}(x(\cdot), R_1) \le C\sqrt{\varepsilon + \omega_J(\varepsilon)}$. *Furthermore, $C(\infty) < \infty$ when $L < 0$.*

Proof. Fix $h > 0$ and consider the uniform grid of I with length h. Replacing δ in (2.3) by 0, we see that there exists a constant C_1 such that $\operatorname{dist}(x(\cdot), R_2^{\Delta}) \le C_1\sqrt{h + \varepsilon + \omega_J(h+\varepsilon)}$. By Theorem 2.2, there exists a constant C_2 such that

$$D_H(R_2^{\Delta}, R_1) \le C_2\sqrt{h + \omega_J(h)}$$

for each $h > 0$. Since $\operatorname{dist}(x(\cdot), R_1) \le \operatorname{dist}(x(\cdot), R_2^{\Delta}) + D_H(R_2^{\Delta}, R_1)$, we can finish the proof by letting $h \to 0$. □

After some standard calculations one can also prove the following extension of Corollary 2.3.

If $x(\cdot)$ is a solution of

$$\dot{x}(t) \in \overline{co}\ F(x(t) + \varepsilon\mathbb{B}),\ x(0) = x_0,$$

then there exists a constant $C = C(T)$ such that $\operatorname{dist}(x(\cdot), R_1) \le C\sqrt{\varepsilon + \omega_J(\varepsilon)}$. Furthermore, $C(\infty) < \infty$ when $L < 0$.

COROLLARY 2.4. *Under the conditions of Theorem* 2.2, $D_H(A(T, x_0), A(T, y_0)) \le e^{LT}|x_0 - y_0|$.

The proof is standard (one can follow the proofs of the previous corollary and the second part of Theorem 2.2 (see, e.g., [7, 8])) and is therefore omitted.

If $F(\cdot)$ is compact-valued and USC, then the reachable set $A(t, x_0)$ is also compact for every $t > 0$. When F is only weakly compact-valued or UHC, it is not clear if $A(t, x_0)$ is always closed. Clearly, the solution set of (1.1) is $C([0,T], H)$ closed.

The following result may be viewed as a "robustness" property of the OSL property.

LEMMA 2.5. *Let $F, G_n : E \rightrightarrows E$ be bounded by N and suppose that for each $M > 0$,*

$$\lim_{n\to\infty} D_H(Graph_{M\mathbb{B}}\ F, Graph_{M\mathbb{B}}\ G_n) = 0.$$

If each G_n is OSL with constant L (not depending on n), then F is also OSL with the same constant L.

Proof. Given $\delta > 0$, let $D_H(\mathrm{Graph}_{M\mathbb{B}}\ F, \mathrm{Graph}_{M\mathbb{B}}\ G) < \delta$, and let G be OSL with a constant L. If $x, y \in M\mathbb{B}$, then there exist $l_x, l_y \in \delta\mathbb{B}$ such that $\sigma(J(x-y), F(x)) < \sigma(J(x-y), G(x+l_x)) + \delta$ and $\sigma(J(x-y), F(y)) > \sigma(J(x-y), G(y+l_y)) - \delta$. Consequently,

$$\begin{aligned}
&\sigma(J(x-y), F(x)) - \sigma(J(x-y), F(y)) \\
&\le \sigma(J(x-y), G(x+l_x)) - \sigma(J(x-y), G(y+l_y)) + 2\delta \\
&\quad \le \sigma(J(x+l_x-y-l_y), G(x+l_x)) - \sigma(J(x+l_x-y-l_y), G(y+l_y)) \\
&+ 2N\Big|J(x-y) - J(x+l_x-y-l_y)\Big| + 2\delta \\
&\le L|x+l_x-y-l_y|^2 + 4N\omega_J(\delta) + 2\delta \\
&\quad \le L|x-y|^2 + 4\delta|L|(2M+\delta) + 4N\omega_J(\delta) + 2\delta.
\end{aligned}$$

Since $\lim_{\delta\to 0}(4\delta|L|(M+\delta) + 4N\omega_J(\delta) + 2\delta) = 0$, one has $\sigma(J(x-y), F(x)) - \sigma(J(x-y), F(y)) \le L|x-y|^2$. ∎

Denote $h_i = t_i - t_{i-1}$. From Lemma 2.1 we know that there exist constants M and N (depending on L and T) such that $|x(t)| \le M$ and $|F(x(t)+\mathbb{B})| \le N$ for every solution $x(\cdot)$ of (1.1), when $F(x)$ is replaced by $F(x(t)+\mathbb{B})$. Thus for $h_\Delta N < 1$, every solution of (1.2) is a solution of $\dot{x}(t) \in F(t, x(t)+\mathbb{B})$.

To estimate the distance between the solution sets of (1.1) and (1.2), for a given partition Δ of $[0,T]$, we fix x_0 and set $d_k = D_H(A(t_k, x_0), A^\Delta(k))$.

COROLLARY 2.6. *Let $F(\cdot)$ be OSL, bounded on bounded sets, UHC, and convex weakly compact-valued. If $h_k \le \min\{1, \frac{1}{N}\}$, then there exists a constant C such that for every $1 \le k \le n$,*

$$d_{k+1}^2 \le e^{2Lh_k} d_k^2 + Ch_k(h_k + \omega_J(h_k)).$$

Proof. We will modify the proof of Lemma 3.13 in [9]. Let $x(\cdot)$ be a solution of (1.1) and let $\varepsilon > 0$ be given. We first find a solution $y(\cdot)$ of (1.2) such that $|x(t_k) - y(t_k)| < \mathrm{dist}(x(t_k), A^\Delta(k)) + \varepsilon$. Since $F(\cdot)$ is OSL and convex weakly compact-valued, there exists a strongly measurable selection $f(t) \in F(y(t_k))$ on $[t_k, t_{k+1}]$ such that

$$\langle J(x(t) - y(t_k)), \dot{x}(t) - f(t)\rangle \le L|x(t) - y(t_k)|^2.$$

For $t \in [t_k, t_{k+1}]$, define $y(t) = y(t_k) + \int_{t_k}^t f(\tau)\, d\tau$. We have

$$\begin{aligned}
&\langle J(x(t) - y(t)), \dot{x}(t) - \dot{y}(t)\rangle \le L|x(t) - y(t_k)|^2 \\
&+ |J(x(t) - y(t)) - J(x(t) - y(t_k))| \cdot |\dot{x}(t) - \dot{y}(t)| \\
&\le L|x(t) - y(t_k)|^2 + 2N\omega_J(N(t - t_k)).
\end{aligned}$$

Let $r(t) = x(t) - y(t)$. Then $\langle J(r(t)), \dot{r}(t)\rangle \le L|x(t) - y(t_k)|^2 + 2N^2\omega_J(t - t_k)$. Furthermore, by (1.3) and Lemma 2.1,

$$|x(t) - y(t_k)|^2 \le |x(t) - y(t)|^2 + 4MN(t - t_k).$$

Consequently, $\langle J(r(t)), \dot{r}(t)\rangle \le L|r(t)|^2 + 2N^2\omega_J(t - t_k) + 4MN|L|(t - t_k)$. Hence $\frac{d}{dt}|r(t)|^2 \le 2L|r(t)|^2 + C\left(\omega_J(t - t_k) + (t - t_k)\right)$. By standard calculations one can show that $|r(t_{k+1})|^2 \le e^{2Lh_k}|r(t_k)|^2 + O(h_k(h_k + \omega_J(h_k)))$. Therefore we derive

$$(\mathrm{dist}(x(t_{k+1}), A^\Delta(k+1)))^2 \le e^{2Lh_k}(\mathrm{dist}(x(t_k), A^\Delta(k)) + \varepsilon)^2 + O\left(h_k(h_k + \omega_J(h_k))\right).$$

Since $\varepsilon > 0$ is arbitrary, one may replace it by 0. The fact that for every solution $y(\cdot)$ of (1.2),

$$(\mathrm{dist}(y(t_{k+1}), A(t_{k+1})))^2 \le e^{2Lh_k}(\mathrm{dist}(y(t_k), A(t_k)))^2 + C(h_k(h_k + \omega_J(h_k)))$$

can be proved as in Theorem 2.2. Thus $d_{k+1}^2 \le e^{2Lh_k}d_k^2 + Ch_k(h_k + \omega_J(h_k))$. ∎

Using the discrete Gronwall inequality, we obtain another corollary.

COROLLARY 2.7. *Under the assumptions of the previous corollary, for a uniform partition of the interval* $[0, T]$ *with step* h *and* $T = nh$,

$$D_H(A(T, x_0), A^\Delta(n, y_0)) \le e^{LT}|y_0 - x_0| + C\sqrt{h + \omega_J(h)},$$

where C *does not depend on* T *if* $L < 0$.

3. The Euler method. We denote by $(I + hF)(X_0) = \bigcup_{x\in X_0}\{x + hF(x)\}$ an Euler step for the inclusion (1.1).

Higher powers of $I + hF$ are defined by

$$(I + hF)^{n+1}(X_0) = (I + hF)\left[\left(I + hF\right)^n(X_0)\right].$$

Consider the Euler discretization of (1.1) for a given partition Δ of $[0, T]$:

$$x(t) = x_i + (t - t_i)f_i, \tag{3.1}$$

where $f_i \in F(x_i),\ x_i = x(t_i),\ x(0) = x_0,\ t \in [t_i, t_{i+1}],\ i = 0, 1, \ldots$.

Denote the reachable set of (3.1) at time $t \in [0, T]$ by $Reach^\Delta(t, x_0)$ or by $Reach^\Delta(t)$ for short.

The following corollary extends the exponential formula of Wolenski [20] and its OSL extension in $\mathbb{R}^n$ [10] to the case of UHC differential inclusions in a Banach space.

COROLLARY 3.1 (exponential formula). *Under the conditions of Theorem* 2.2, *for every* $t \in [0, T]$ *we have*

$$A(T, x_0) = \lim_{n\to\infty}\left(I + \frac{T}{n}F\right)^n(x_0) := e^{TF}x_0.$$

Proof. For the uniform partition $t_i = \frac{iT}{n},\ i = 0, 1, \ldots, n-1$, with $h = \frac{T}{n}$, consider the Euler polygons (3.1).

Since $F(\cdot)$ is convex weakly compact-valued, the reachable set of (3.1) coincides at the points t_k with the reachable set of (1.2), i.e., $A^\Delta(t_k, x_0) = Reach^\Delta(t_k, x_0) = (I + hF)^k(x_0)$. It follows from Corollaries 2.6 and 2.7 that

$$D_H\left(A(T, x_0), \left(I + \frac{T}{n}F\right)^n(x_0)\right) \le C\sqrt{\frac{1}{n} + \omega\left(\frac{1}{n}\right)}.$$

The proof is therefore complete. ∎

Next we assume that $L < 0$ and study the asymptotic behavior of the reachable set when $T \to \infty$.

The following theorem extends Theorems 3.1 and 3.5 of [9].

THEOREM 3.2. *Suppose that $F(\cdot)$ satisfies all the assumptions of Theorem* 2.2 *with $L < 0$. Then there exists a closed and bounded set A^∞ such that $A^\infty = \lim_{t\to\infty} \overline{A(t, x_0)}$. Moreover,*

(i) *A^∞ is attracting the reachable sets; i.e., for bounded $B \neq A^\infty$, if $T > 0$, then $D_H(A(T, B), A^\infty) < D_H(B, A^\infty)$;*

(ii) *$\overline{A(T, A^\infty)} = A^\infty$ for every $T > 0$;*

(iii) *A^∞ does not depend on x_0;*

(iv) *A^∞ is a strongly invariant set for* (1.1)*; i.e., every solution $x(\cdot)$ of* (1.1) *with $x(0) \in A^\infty$ satisfies $x(t) \in A^\infty$ for every $t > 0$.*

Proof. We will follow the proofs of Theorems 3.1 and 3.5 of [9]. Let $0 < t < s$. Since for every bounded B, the sets $A(t, B)$ and $A(s, B)$ are bounded subsets of $M\mathbb{B}$, by Corollary 2.4, $D_H(A(t, B), A(s, B)) \le D_H(B, A((s - t), B))e^{L(s-t)} \le 2Me^{L(s-t)}$. Thus the net $\{A(t, B)\}_{t\ge 0}$ is a Cauchy net. Therefore, $\{\overline{A(t, B)}\}_{t\ge 0}$ is also a Cauchy net. Since the set of closed and bounded subsets of E equipped with the Hausdorff metric is a complete metric space, the limit A^∞ exists. By Corollary 2.4, this limit does not depend on the initial condition x_0 and satisfies (i) because $L < 0$. Moreover, for a given $T > 0$, the multivalued map $B \to \overline{A(T, B)}$ is a set-valued contraction. Hence there exists a unique nonempty closed set D_T with $\overline{A(T, D_T)} = D_T$. As in [9], it is easy to see that $D_T = A^\infty$. Hence $\overline{A(T, A^\infty)} = A^\infty$ for every $T \ge 0$. Thus A^∞ is a strongly invariant set. The rest of the proof is obvious. ∎

COROLLARY 3.3. *Under the conditions of Theorem* 3.2*, $A^\infty \subset B$ for every other strongly invariant set B of* (1.1)*.*

Proof. By (i) of Theorem 3.2, we see that $\lim_{t\to\infty} \overline{A(t, B)} = A^\infty \subset B$, because B is a strongly invariant set. ∎

We call A^∞ the *fixed set* (for (1.1)).

The next theorem shows that the fixed set may be approximated by the reachable set $Reach^\Delta(t, x_0)$ of (3.1) for special time step sequences and may be regarded as an infinite time exponential formula (cf. Corollary 3.15 of [9]) in the setting of Banach spaces.

THEOREM 3.4 (fixed set iterations). *Under the conditions of Theorem* 3.2*, there exists a partition Δ of $[0, \infty)$ such that the reachable set $Reach^\Delta(t, x_0)$ of* (3.1) *satisfies $\lim_{t\to\infty} D_H(Reach^\Delta(t, x_0), A^\infty) = 0$.*

Proof. By (ii) of Theorem 3.2, $A^\infty = \overline{A(t, A^\infty)}$. Furthermore, by Corollary 2.4 we know that

$$D_H(A(T, A_0), A(T, A^\infty)) \le e^{LT} D_H(A_0, A^\infty). \tag{3.2}$$

One may suppose that the constant $C(T)$ in Theorem 2.2 is increasing; i.e., $C(T) \le C(\infty)$.

We let $A_0 = \{x_0\}$, find $T > 0$ such that $e^{LT} < \frac{1}{4}$, and set $i := 0$. Now we proceed as follows.

Step A. Set $K_i = D_H(A_i, A^\infty)$. Hence $D_H(A^\infty, A(T, A_i)) < \frac{K_i}{4}$. Let h_i be so small that for $h_i := \frac{T}{k_i}$ one has $C(\infty)\sqrt{h_i + \omega_J(h_i)} < \frac{K_i}{4}$. Let Δ_i be the uniform partition of $[iT, (i + 1)T]$ corresponding to h_i. It follows from Corollary 3.1 that

$$D_H(A(T, A_i), Reach^{\Delta_i}(T, A_i)) \le C(\infty)\sqrt{h_i + \omega_J(h_i)} < \frac{K_i}{4}.$$

Let $A_{i+1} := Reach^{\Delta_i}(T, A_i)$. Consequently,

$$D_H(A_{i+1}, A^\infty) \le D_H(A_{i+1}, A(T, A_i)) + D_H(A(T, A_i), A^\infty) < 2\frac{K_i}{4} = \frac{K_i}{2}. \tag{3.3}$$

End of Step A.

Since the system is autonomous, one can start from A_1 and repeat Step A with A_0 replaced by A_1 and K_0 replaced by K_1 $\big($hence $K_1 < \frac{K_0}{4}\big)$. We find $h_1 < h_0$ and A_2 such that $D_H(A_2, A^\infty) < \frac{K_1}{2} < \frac{K_0}{4}$. Notice that A_2 is the reachable set of (3.1) for the time $2T$, where the partition Δ of the interval $[0, 2T]$ has step h_0 on $[0, T]$ and step h_1 on $[T, 2T]$.

Now we repeat Step A n times and derive a sequence $\{A_n\}_{n=1}^\infty$ where A_n is the reachable set of (3.1) for the time $nT := \sum_{i=0}^{n-1} T$ and $D_H(A_n, A^\infty) < \frac{K_0}{2^n}$. It is easy to see that $\lim_{n\to\infty} nT = \infty$ and $\lim_{n\to\infty} \overline{A_n} = A^\infty$, which yields the assertion of the theorem. □

Note that the choice of the time steps above depends on the modulus ω and therefore cannot be the same as in the corresponding result of [9].

4. Controllability of other-sided Lipschitz systems. We will call the multimap $F : E \rightrightarrows E$ other-sided (backward-sided) Lipschitz (BSL) when

$$\sigma(J(x-y), F(x)) - \sigma(J(x-y), F(y)) \ge L|x-y|^2 \text{ for all } x, y \in E.$$

It is easy to see that $F(\cdot)$ is BSL if and only if the mapping $-F(\cdot)$ is OSL.

Consider the differential inclusion

$$\dot{x}(t) \in F(x),\ x(0) = x_0 \in K. \tag{4.1}$$

If $F(\cdot)$ is BSL, it is not possible to prove either existence of solutions or qualitative properties. However, changing the time direction by setting $t = -s$, we obtain an OSL differential inclusion. Let $F(\cdot)$ be BSL with constant $L > 0$. In this case $-F(\cdot)$ is OSL with constant $-L$. Assume that $F(\cdot)$ has nonempty, convex, and weakly compact values and is bounded on bounded sets. By Theorem 3.2, there exists a backward fixed set $A^{-\infty} = \lim_{t\to-\infty} \overline{A(t, x_0)}$ such that

(i) $A^{-\infty}$ is a backward attractor of the reachable sets; that is, if $B \ne A^{-\infty}$ is bounded and $T < 0$, then $D_H(A(T, B), A^\infty) < D_H(B, A^\infty)$;

(ii) $\overline{A(T, A^{-\infty})} = A^{-\infty}$ for every $T < 0$;

(iii) $A^{-\infty}$ does not depend on x_0;

(iv) $A^{-\infty}$ is a strongly backward invariant set for (1.1); i.e., every solution $x(\cdot)$ of (1.1) with $x_0 \in A^{-\infty}$ satisfies $x(t) \in A^{-\infty}$ for every $t < 0$.

Note that the above assertions are those of Theorem 3.2 rewritten in the backward time direction. As a corollary we derive the following criterion for complete controllability.

THEOREM 4.1. *Let $F(\cdot)$ be UHC, with nonempty, convex, and weakly compact values, bounded on bounded sets, and BSL with positive constant L.*

(i) *If K is an open set such that $K \bigcap A^{-\infty} \ne \emptyset$, then for every $z \in E$, there exist $x_0 \in K$ and a finite time $T(z)$ such that z belongs to the reachable set of* (4.1) *at time $T(z)$; i.e.,* (4.1) *is completely quasi-controllable.*

(ii) *If K is an open set such that $K \bigcap A^{-\infty} = \emptyset$, then there exists $z \in E$ which cannot be reached in finite time from K; i.e.,* (4.1) *is not completely quasi-controllable.*

Finally, we present an example of a Hilbert space differential inclusion to which our results can be applied.

EXAMPLE 4.1. *Let $E = l^2$ be the Hilbert space of all square summable real sequences.*

Consider the following system:

$$
\begin{aligned}
&\dot{x}_0(t) \in -x_0(t) + [-1,1],\ x_0(0) = 0,\\
&\dot{x}_1(t) \in -x_1(t) - \frac{\sqrt[5]{x_1^3}}{\sqrt[3]{1}} + \frac{1}{2}[-|x_0(t)|, |x_0(t)|],\ x_1(0) = 0,\\
&\qquad \dots \qquad (4.2)\\
&\dot{x}_n(t) \in -x_n(t) - \frac{\sqrt[5]{x_n^3}}{\sqrt[3]{n}} + \frac{1}{2}[-|x_{n-1}(t)|, |x_{n-1}(t)|],\ x_n(0) = 0,\\
&\dots .
\end{aligned}
$$

It is not difficult to see that the right-hand side F of (4.2) *is OSL with a constant $L = -\frac{1}{2}$. We claim that F maps l^2 into l^2. If $\vec{x} \in l^2$, then obviously $-\vec{x} \in l^2$, too. Also, if $\vec{x} \in l^2$ with $\vec{x} = (x_0, x_1, \dots, x_n, \dots)$, then $\vec{y} = (0, |x_0|, |x_1|, \dots, |x_n|, \dots) \in l^2$. Thus we have only to check that if $\vec{x} = (x_0, x_1, \dots, x_n, \dots) \in l^2$, then $\sum_{n=0}^{\infty} \frac{\sqrt[5]{x_n^6}}{\sqrt[3]{n^2}}$ converges. But this is a trivial consequence of the Cauchy–Schwarz inequality ($\langle \vec{x}, \vec{y} \rangle \le |\vec{x}||\vec{y}|$) and the fact that the series $\sum_{n=1}^{\infty} \sqrt[5]{x_n^{12}}$ and $\sum_{n=1}^{\infty} \sqrt[3]{n^{-4}}$ converge.*

Note that in this example $F(\cdot)$ is continuous (not Lipschitz) with convex (strongly) compact values. The existence of solutions to the system (4.2) *is proved in* [4]. *This existence also follows from Theorem* 2.2.

Acknowledgments. All the authors thank the referees for many valuable comments and suggestions.

REFERENCES

[1] J.-P. AUBIN AND A. CELLINA, *Differential Inclusions*, Springer-Verlag, Berlin, 1984.

[2] A. BULGAKOV AND V. SKOMODOV, *Approximation of differential inclusions*, Math. Sbornik, 193 (2002), pp. 35–52 (in Russian).

[3] K. DEIMLING, *Multivalued Differential Equations*, De Grujter, Berlin, 1992.

[4] T. DONCHEV, *Semicontinuous differential inclusions*, Rend. Sem. Mat. Univ. Padova, 101 (1999), pp. 147–160.

[5] T. DONCHEV, *Differential inclusions in uniformly convex spaces. Baire category approach* II, Ann. Şti. Univ. Ovidius Constanţa, 7 (1999), pp. 67–76.

[6] T. DONCHEV, *Approximation of lower semicontinuous differential inclusions*, Numer. Funct. Anal. Optim., 22 (2001), pp. 55–67.

[7] T. DONCHEV AND E. FARKHI, *Stability and Euler approximation of one-sided Lipschitz differential inclusions*, SIAM J. Control Optim., 36 (1998), pp. 780–796.

[8] T. DONCHEV AND E. FARKHI, *Euler approximation of discontinuous one-sided Lipschitz convex differential inclusions*, in Calculus of Variations and Differential Equations, A. Ioffe, S. Reich, and I. Shafrir, eds., Chapman & Hall/CRC, Boca Raton, New York, 1999, pp. 101–118.

[9] T. DONCHEV, E. FARKHI, AND S. REICH, *Fixed set iterations for relaxed Lipschitz multimaps*, Nonlinear Anal., 53 (2003), pp. 997–1015.

[10] T. DONCHEV, E. FARKHI, AND P. WOLENSKI, *Characterizations of reachable sets for a class of differential inclusions*, Funct. Differ. Equ., 10 (2003), pp. 473–483.

[11] A. DONTCHEV AND E. FARKHI, *Error estimates for discretized differential inclusions*, Computing, 41 (1989), pp. 349–358.

[12] G. GRAMMEL, *Towards fully discretized differential inclusions*, Set-Valued Anal., 11 (2003), pp. 1–8.

[13] G. GRAMMEL, *On the Time Discretization of Singularly Perturbed Systems*, Lecture Notes in Comput. Sci. 3743, Springer-Verlag, New York, 2006, pp. 297–304.

[14] S. HU AND N. S. PAPAGEORGIOU, *Handbook of Multivalued Analysis, Vol.* I, *Theory*, Math. Appl. 419, Kluwer, Dordrecht, The Netherlands, 1997.

[15] S. Hu and N. S. Papageorgiou, *Handbook of Multivalued Analysis, Vol.* II, *Applications*, Math. Appl. 500, Kluwer, Dordrecht, The Netherlands, 2000.
[16] V. Lakshmikantham and S. Leela, *Nonlinear Differential Equations in Abstract Spaces*, Pergamon, Oxford, UK, 1981.
[17] F. Lempio and V. Veliov, *Discrete approximations of differential inclusions*, Bayreuth. Math. Schr., 54 (1998), pp. 149–232.
[18] B. S. Mordukhovich, *Discrete approximations and refined Euler–Lagrange conditions for nonconvex differential inclusions*, SIAM J. Control Optim., 33 (1995), pp. 882–915.
[19] I. I. Vrabie, *A Nagumo type viability theorem*, An. Ştiinţ. Univ. Al. I. Cuza Iaşi. Mat. (N.S.), 51 (2005), pp. 293–308.
[20] P. R. Wolenski, *The exponential formula for the reachable set of a Lipschitz differential inclusion*, SIAM J. Control Optim., 28 (1990), pp. 1148–1161.

SIAM J. OPTIM.
Vol. 18, No. 3, pp. 907–913

THE VECTOR-VALUED VARIATIONAL PRINCIPLE IN BANACH SPACES ORDERED BY CONES WITH NONEMPTY INTERIORS*

EWA M. BEDNARCZUK† AND MACIEJ J. PRZYBYŁA‡

Abstract. We prove sharpness of efficient solutions x_k to vector optimization problems resulting from Ekeland vector variational principles. We achieve this by sharpening some of the existing vector variational principles and showing that x_k remains efficient not only for perturbations in the direction k but also for other directions of perturbations.

Key words. vector-valued variational principle, Bishop–Phelps cone, sharp solutions

AMS subject classifications. 58E30, 58E17, 65K10

DOI. 10.1137/060658989

1. Introduction. In its classical formulation, the Ekeland variational principle [3] states that in a vicinity of an ε-solution to the minimization problem of a lower semicontinuous (*lsc*) bounded below real-valued function f defined on a complete metric space, one can always find a strict solution to a minimization problem with slightly perturbed function f. There exist many generalizations of this theorem, to metric spaces, e.g., [2, 8, 13], to locally convex topological spaces and to general topological spaces [4]. A parametrized version of Ekeland's variational principle was proved in [7].

In recent years variational principles for vector-valued functions taking values in partially ordered spaces have been studied by several authors [5, 6, 10, 11, 12, 16].

The following vector variational principle was proved in [5] and [16].

THEOREM 1.1 (see Theorem 8 in [5]). *Let X and Z be real Banach spaces and Z be partially ordered by a closed convex pointed cone $\mathcal{K}$. An element $+\infty \notin Z$ is such that $z \leq +\infty$ for all $z \in Z$. Let $f\colon X \to Z \cup \{+\infty\}$ be a quasi-lsc, bounded below proper function, $\bar{x} \in D(f)$, and let $k \in \mathcal{K} \setminus \{0\}$. If $f(X) \cap (f(\bar{x}) - \varepsilon k - \mathcal{K} \setminus \{0\}) = \emptyset$ for some $\varepsilon > 0$, then for every $\lambda > 0$ there exists x_k such that*

(i) $f(x_k) \leq f(\bar{x})$,
(ii) $\|\bar{x} - x_k\| \leq \lambda$,
(iii) $x_k \in E(f_{(x_k,k)})$, *where* $f_{(x_k,k)}(x) := f(x) + \frac{\varepsilon}{\lambda}\|x - x_k\|k$.

The following theorem [9, 10] relaxes the requirement for f to be bounded below, strengthens assertion (iii), and puts less stringent assumptions on X, Y, and $\mathcal{K}$.

THEOREM 1.2 (see Corollary 3.10.14 in [9]). *Let (X, d) be a complete metric space and Y be a separated locally convex space ordered by a closed cone $\mathcal{K}$. An element $+\infty \notin Y$ is such that $y \leq +\infty$ for all $y \in Y$. Let $f\colon X \to Y \cup \{+\infty\}$ be a proper function, $\bar{x} \in D(f)$, and let $k \in \mathcal{K} \setminus (-\mathcal{K})$. Suppose that for every $r \in \mathbb{R}$ the set $\{x \in X \mid f(x) \leq f(\bar{x}) + rk\}$ is closed. If $f(X) \cap (f(\bar{x}) - \varepsilon k - \mathcal{K} \setminus \{0\}) = \emptyset$ for some $\varepsilon > 0$, then for every $\lambda > 0$ there exists x_k such that*

(i) $f(x_k) + \lambda^{-1}\varepsilon d(x_k, \bar{x})k \leq f(\bar{x})$, $d(x_k, \bar{x}) \leq \lambda$,

*Received by the editors May 4, 2006; accepted for publication (in revised form) March 21, 2007; published electronically October 4, 2007.
http://www.siam.org/journals/siopt/18-3/65898.html
†Systems Research Institute, Polish Academy of Sciences, ul. Newelska 6, 01-447 Warszawa, Poland (Ewa.Bednarczuk@ibspan.waw.pl).
‡Ph.D. Studies, Systems Research Institute, Polish Academy of Sciences, ul. Newelska 6, 01-447 Warszawa, Poland (M.Przybyla@ibspan.waw.pl).

(ii) *if* $f(x) + \lambda^{-1}\varepsilon d(x_k, x)k \leq f(x_k)$, *then* $x = x_k$.

In [6], Finet, Quarta, and Troestler introduced the notion of the order lower semicontinuity and deduced the vector-valued Ekeland variational principle for an order lower semicontinuous bounded below function f directly from the Deville–Godefroy–Zizler perturbed minimization principle (Corollary 31 of [6]). In [11] and [12], a relation between nuclearity of cones in product spaces and Ekeland's variational principle is considered.

Sharp efficiency can be viewed as one of the weakest types of proper efficiency and plays a crucial role in investigating stability to perturbed vector optimization problems (cf. [1]).

In the present paper we prove that x_k is a sharp efficient solution (in the sense of [1]) to function $f_{(x_k,k)}(x) = f(x) + \frac{\varepsilon}{\lambda}\|x - x_k\|$. Moreover, we introduce the new concept of ε-solutions to vector optimization problems with respect to Bishop–Phelps cones. By using this concept we derive another form of vector variational principle which gives an efficient solution x_0 to functions $f_{(x_0,k)}$ for all $k \in \operatorname{int} \mathcal{K}_\alpha$ with $\varphi(k) - \alpha\|k\|$ sufficiently large and we prove the sharpness of x_0.

The organization of the paper is as follows. In section 3 we introduce a new concept of ε-efficient solutions for f taking values in spaces ordered by Bishop–Phelps cones $\mathcal{K}_\alpha$. This concept strengthens some of the existing definitions of ε-efficiency and allows us to prove the vector-valued Ekeland variational principle for any $k \in \operatorname{int} \mathcal{K}_\alpha$ and to deduce the sharpness of solutions x_k. In section 4 we prove our main results, which are in Theorems 4.1 and 4.2.

2. Preliminaries. Let $X = (X, d)$ be a complete metric space. Let Z be a separated locally convex space ordered by the relation $\leq$ ($\geq$) induced by a closed cone $\mathcal{K}$ in the usual way; i.e., for any $z, y \in Z$ we put $z \leq y$ ($y \geq z$) iff $y - z \in \mathcal{K}$.

Let $\bar{Z} := Z \cup \{+\infty\}$, where $+\infty \notin Z$ is such that $z \leq +\infty$ for any $z \in Z$ and let f be a function from X to $\bar{Z}$. The domain $D(f)$ of f is defined as $D(f) := \{x \in X \mid f(x) \neq +\infty\}$ and f is called *proper* if $D(f) \neq \emptyset$. Consider the vector minimization problem

$$(P) \qquad \begin{array}{l} f(x) \to_{\mathcal{K}} \min \\ \text{subject to } x \in X. \end{array}$$

A point $\bar{x} \in X$ is an *efficient solution* to (P) if

$$(f(\bar{x}) - \mathcal{K}) \cap f(X) \subset f(\bar{x}) + \mathcal{K}.$$

We denote by $E(f)$ the set of all efficient solutions to (P).

A function $f\colon X \to \bar{Z}$ is quasi-*lsc* at $x_0 \in X$ [5], if for all $b \in Z$ such that $b \ngeq f(x_0)$ there exists a neighborhood U of x_0 in X such that $b \ngeq f(x)$ for all $x \in U$. f is quasi-*lsc* if f is quasi-*lsc* at each point of X. f is quasi-*lsc* iff for all $b \in Z$ the sets $\{x \in X \mid f(x) \leq b\}$ are closed in X.

A function $f\colon X \to \bar{Z}$ is *lsc* at $x_0 \in D(f)$ if for each neighborhood W of $f(x_0)$ there exists a neighborhood V of x_0 such that $f(V) \subset W + (\mathcal{K} \cup +\infty)$. If f is *lsc* at x_0, then f is quasi-*lsc* at x_0.

Definition 2.1 (see [16]). *For given* $\varepsilon > 0$ *and* $k \in \mathcal{K} \setminus \{0\}$, *a point* $\bar{x} \in D(f)$ *is* ε*-approximately efficient in the direction* k *if* $(f(\bar{x}) - \varepsilon k - \mathcal{K} \setminus \{0\}) \cap f(X) = \emptyset$.

The notion of ε-approximately efficient points is essential in proving Theorems 1.2 and 1.1 (see also [5, 9, 10, 16]).

Let Z^* be the dual space of all continuous linear functionals defined on Z and $\mathcal{K}^* := \{z^* \in Z^* \mid \forall\, z \in \mathcal{K}\ z^*(z) \geq 0\}$ the dual cone of $\mathcal{K}$. Since $\mathcal{K}$ is pointed, for any

$k \in \mathcal{K} \setminus \{0\}$, by the Hahn–Banach theorem there exists $k^* \in \mathcal{K}^*$ such that $k^*(k) = 1$. The functional k^* is order-preserving. Let us recall that a function $f\colon X \to Z$ is said to be *bounded below* if there exists $z \in Z$ such that $z \le f(x)$ for all $x \in X$. If f is bounded below, then $k^* \circ f$ is bounded below on X and $k^*(z) \le (k^* \circ f)(x)$ for all $x \in X$ and any $k^* \in \mathcal{K}^*$; on the other hand the function $k^* \circ f$ may not be *lsc*.

3. The variational principle for Bishop–Phelps cones. Let $0 < \alpha < 1$ be given. For an arbitrary $\varphi \in Z^*$, $\|\varphi\| = 1$ we define

$$\mathcal{K}_\alpha := \{z \in Z \mid \varphi(z) \ge \alpha\|z\|\}.$$

Let us notice that $\mathcal{K}_\alpha$ is a closed convex pointed cone with nonempty interior $\operatorname{int}\mathcal{K}_\alpha = \{z \in Z \mid \varphi(z) > \alpha\|z\|\}$ and a bounded base $\Theta = \{k \in \mathcal{K}_\alpha \mid \varphi(k) = 1\}$. The cone $\mathcal{K}_\alpha$ is called a Bishop–Phelps cone [14, 15].

In what follows we assume that Z is a real Banach space partially ordered by a cone $\mathcal{K}_\alpha$.

In what follows we use the functions $g\colon \bar{Z} \to \bar{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$ defined by

$$g(z) := \begin{cases} \varphi(z) - \alpha\|z\| & \text{for } z \in Z, \\ +\infty & \text{for } z = +\infty, \end{cases}$$

and the function $v\colon X \to \bar{\mathbb{R}}$ defined by $v := g \circ f$:

$$v(x) := \varphi(f(x)) - \alpha\|f(x)\|. \tag{3.1}$$

The function g is order-preserving since for any $z_1 \le z_2$ we have $g(z_1) \le g(z_2)$. In consequence, if f is bounded below on X, i.e., $z \le f(x)$ for any $x \in X$, the function v is also bounded below on X and $g(z) \le v(x)$ for all $x \in X$. The converse is not true.

Moreover,

$$S(v) := \operatorname{argmin}\{v(x) \ : \ x \in X\} \subset WE(f),$$

where $WE(f)$ denotes the set of all weakly efficient solutions to (P), i.e., $\bar{x} \in WE(f)$ if $(f(\bar{x}) - \operatorname{int}\mathcal{K}_\alpha) \cap f(X) = \emptyset$. To see this, take any $\bar{x} \notin WE(f)$. There exists $x \in X$ such that $f(\bar{x}) - f(x) \in \operatorname{int}\mathcal{K}_\alpha$ and, consequently,

$$\varphi(f(\bar{x})) - \alpha\|f(\bar{x})\| > \varphi(f(x)) - \alpha\|f(x)\|$$

which proves that $\bar{x} \notin S(v)$. Since g is not strongly increasing (i.e., $z_1 \ge z_2$ and $z_1 \ne z_2$ does not imply that $g(z_1) > g(z_2)$), we cannot prove that $S(v) \subset E(f)$.

We introduce the following notion of ε-approximate efficiency with respect to Bishop–Phelps cones.

Definition 3.1. *Let $\varepsilon > 0$ be given and $0 < \alpha < 1$. A point $\bar{x} \in X$ is an ε-efficient solution with respect to $\mathcal{K}_\alpha$ if*

$$v(\bar{x}) < \inf_{x \in X} v(x) + \varepsilon,$$

where the function v is defined by (3.1).

This means that $\bar{x}$ is an ε-solution of the scalar-valued function v.

The next lemma establishes the relationship between the ε-efficiency with respect to $\mathcal{K}_\alpha$ and the ε-approximate efficiency in the sense of Definition 2.1.

Lemma 3.2. *An ε-efficient solution $\bar{x} \in X$ with respect to $\mathcal{K}_\alpha$ is ε_0-approximately efficient in any direction $k \in \operatorname{int}\mathcal{K}_\alpha$ with $\varepsilon_0 \ge \varepsilon/g(k)$.*

Proof. Suppose on the contrary that $\bar{x} \in X$ is not ε_0-approximately efficient in the direction $k \in \operatorname{int} \mathcal{K}_\alpha$. Therefore, by the definition of the ε_0-approximate efficiency there exists $x \in X$ which satisfies

$$\varphi(f(\bar{x}) - (f(x) + \varepsilon_0 k)) \geq \alpha \|f(\bar{x}) - (f(x) + \varepsilon_0 k)\|.$$

Hence in view of the definition of v we obtain the inequality

$$v(\bar{x}) \geq v(x) + \varepsilon_0 g(k),$$

which contradicts the fact that $\bar{x}$ is a ε-efficient solution with respect to $\mathcal{K}_\alpha$. This completes the proof. □

Now we are ready to prove the following vector-valued Ekeland variational principle for a quasi-*lsc* function taking values in Banach spaces partially ordered by Bishop–Phelps cones.

THEOREM 3.3. *Let $f\colon X \to Z \cup \{+\infty\}$ be a quasi-lsc, bounded below function such that $D(f) \neq \emptyset$. Let $\varepsilon > 0$ be given and let $\bar{x} \in X$ be an ε-efficient solution with respect to $\mathcal{K}_\alpha$.*

Then for any $k \in \operatorname{int} \mathcal{K}_\alpha$ there exists $x_k \in X$ such that for any $l \geq k$ the following conditions hold:

(i) $f(x_k) \leq f(\bar{x})$,
(ii) $\|\bar{x} - x_k\| \leq 1/g(k)$,
(iii) *for any $x \in X$ the relation $f(x) + \varepsilon\|x - x_k\| l \leq f(x_k)$ implies $x = x_k$.*

Proof. Take any $k \in \operatorname{int} \mathcal{K}_\alpha$. Since $\bar{x}$ is an ε-efficient solution with respect to $\mathcal{K}_\alpha$, by Lemma 3.2 $\bar{x}$ is ε_0-approximately efficient in the direction k and $\varepsilon_0 = \varepsilon/g(k)$. By Theorem 1.2, taking $\lambda = 1/g(k)$, there exists $x_k \in X$ such that

(i) $f(x_k) \leq f(\bar{x})$,
(ii) $\|\bar{x} - x_k\| \leq 1/g(k)$,
(iii) for any $x \in X$ the relation $f(x) + \varepsilon\|x - x_k\| k \leq f(x_k)$ implies $x = x_k$.

Let $l \geq k$. Since $f(x) + \varepsilon\|x - x_k\| k \leq f(x) + \varepsilon\|x - x_k\| l$, then the relation $f(x) + \varepsilon\|x - x_k\| l \leq f(x_k)$ implies $x = x_k$. □

THEOREM 3.4. *Let X and Z be Banach spaces and let $f\colon X \to \bar{Z}$ be an lsc proper function. Let $\varepsilon > 0$ and $\lambda > 0$ be given and let $\bar{x} \in X$ be an ε-efficient solution with respect to $\mathcal{K}_\alpha$.*

There exists $x_0 \in D(f)$ such that for any $k \in \operatorname{int} \mathcal{K}_\alpha$ such that $g(k) \geq \frac{1}{\lambda}$ the following conditions hold:

(i) $v(x_0) + \frac{\varepsilon}{\lambda}\|\bar{x} - x_0\| \leq v(\bar{x})$,
(ii) $\|\bar{x} - x_0\| \leq \lambda$,
(iii) *for all $x \in X$ the relation $f(x) + \varepsilon\|x - x_0\| k \leq f(x_0)$ implies $x = x_0$.*

Proof. As v is a bounded below *lsc* function and $\bar{x}$ is an ε-solution to v, by the Ekeland variational principle there exists $x_0 \in D(f)$ such that $v(x_0) + \frac{\varepsilon}{\lambda}\|\bar{x} - x_0\| \leq v(\bar{x})$, $\|\bar{x} - x_0\| \leq \lambda$, and

$$\forall\ x \in X\ \ v(x) + \frac{\varepsilon}{\lambda}\|x - x_0\| \leq v(x_0) \ \ \Rightarrow \ \ x = x_0.$$

Take any $k \in \mathcal{K}_\alpha$ such that $g(k) \geq \frac{1}{\lambda}$. Since $g(z + \gamma k) \geq g(z) + \gamma g(k)$ for $z \in Z$ and $\gamma \in \mathbb{R}_+$,

$$\forall\ x \in X\ \ g(f(x) + \varepsilon\|x - x_0\| k) \leq g(f(x_0)) \ \ \Rightarrow \ \ x = x_0.$$

Since g is order-preserving, for any $x \in X$ the relation $f(x) + \varepsilon\|x, x_0\| k \leq f(x_0)$ implies $g(f(x) + \varepsilon\|x - x_0\| k) \leq g(f(x_0))$, which proves (iii). □

4. Sharp efficiency. In this section we assume that X and Z are normed spaces. In [1], Bednarczuk investigates the following notion of sharp efficiency: an $\bar{x} \in D(f)$ is a *sharp efficient solution of order* 1 to (P) if there exists a constant $\tau > 0$ such that

$$f(x) \notin f(\bar{x}) + \tau\|x - \bar{x}\|B_Z - \mathcal{K} \quad \text{for } x \in X,\ x \neq \bar{x},$$

where B_Z is open unit ball in Z. We denote the set of all sharp solutions of order 1 to (P) by $St^1(f)$. Let us note that if $\bar{x} \in St^1(f)$, then for all $x \in X$ $f(x) \leq f(\bar{x}) \Rightarrow x = \bar{x}$.

The following theorem sharpens the conclusion (iii) of Theorem 1.2 for cones with nonempty interiors.

THEOREM 4.1. *Let X and Z be normed spaces with Z ordered by a closed pointed cone $\mathcal{K}$, $\operatorname{int}\mathcal{K} \neq \emptyset$. Let $f\colon X \to \bar{Z}$ be a quasi-lsc proper function. Let $\varepsilon > 0$, $\lambda > 0$, and $k \in \operatorname{int}\mathcal{K}$. Let $\bar{x} \in D(f)$ be an ε-approximately efficient point in the direction $k \in \operatorname{int}\mathcal{K}$.*

There exists $x_k \in D(f)$ such that the following conditions hold:

(i) $f(x_k) \leq f(\bar{x})$,

(ii) $\|x_k - \bar{x}\| \leq \lambda$,

(iii) $x_k \in St^1(f_{(x_k,l)})$ *for any* $l \geq 2k$, *where* $f_{(x_k,l)}(x) := f(x) + \frac{\varepsilon}{\lambda}\|x - x_k\|l$.

Proof. By assumption, there exists $x_k \in X$ satisfying the conditions (i), (ii), and (iii) of Theorem 1.2. Since $k \in \operatorname{int}\mathcal{K}$, there exists a number $\tau > 0$ such that $2k - \frac{\lambda\tau}{\varepsilon}b \geq k$ for any $b \in B_Z$. Take any $l \geq 2k$, then $l - \frac{\lambda\tau}{\varepsilon}b \geq k$ for any $b \in B_Z$. By Theorem 1.2(iii),

$$f(x_k) - f(x) - \frac{\varepsilon}{\lambda}\|x - x_k\|\left(l - \frac{\lambda\tau}{\varepsilon}b\right) \notin \mathcal{K} \ \forall\ x \neq x_k.$$

Since the relation holds for any $b \in B_Z$ we have

$$f(x) + \frac{\varepsilon}{\lambda}\|x - x_k\|l \notin f(x_k) + \tau\|x - x_k\|B_Z - \mathcal{K} \ \forall\ x \neq x_k,$$

which proves that $x_k \in St^1(f_{(x_k,l)})$ with the constant τ. □

For $\mathcal{K} = \mathcal{K}_\alpha$ we get the estimation of the distance $\|x_k - \bar{x}\|$ in terms of $g(k)$ and the estimation of constant τ. This is the content of the next theorem.

THEOREM 4.2. *Let $f\colon X \to \bar{Z}$ be a quasi-lsc proper function. Let $\varepsilon > 0$ and let $\bar{x} \in D(f)$ be an ε-efficient solution with respect to $\mathcal{K}_\alpha$.*

For any $k \in \operatorname{int}\mathcal{K}_\alpha$ there exists $x_k \in X$ such that the following conditions are satisfied:

(i) $f(x_k) \leq f(\bar{x})$,

(ii) $\|\bar{x} - x_k\| \leq 1/g(k)$,

(iii) $x_k \in St^1(f_{(x_k,l)})$ *for any* $l \geq 2k$, *where* $f_{(x_k,l)}(x) := f(x) + \varepsilon\|x - x_k\|l$.

Proof. Take any $k \in \operatorname{int}\mathcal{K}$. Let $\tau > 0$ be such that $\tau \leq \varepsilon g(k)/(\|\varphi\| + \alpha)$. We will show that for any $b \in B_Z$ the inequality $2k - \frac{\tau}{\varepsilon}b \geq k$ holds. Indeed

$$\varphi\left(k - \frac{\tau}{\varepsilon}b\right) - \alpha\left\|k - \frac{\tau}{\varepsilon}b\right\| \geq g(k) - \frac{\tau}{\varepsilon}(\varphi(b) + \alpha\|b\|) \geq g(k) - \frac{\tau}{\varepsilon}(\|\varphi\| + \alpha) \geq 0,$$

hence $2k - \frac{\tau}{\varepsilon}b \geq k$ for all $b \in B_Z$. Take any $l \geq 2k$, then $l - \frac{\tau}{\varepsilon}b \geq k$ for any $b \in B_Z$. By Theorem 3.3 (iii), $f(x_k) - f(x) - \varepsilon\|x - x_k\|(l - \frac{\tau}{\varepsilon}b) \notin \mathcal{K}_\alpha$ for $x \neq x_k$. Since the relation holds for any $b \in B_Z$ we have

$$f(x) + \varepsilon\|x - x_k\|l \notin f(x_k) + \tau\|x - x_k\|B_Z - \mathcal{K}_\alpha \ \forall\ x \neq x_k,$$

which completes the proof. □

We close this section by showing two necessary conditions for sharp efficiency. In [6], Finet, Quarta, and Troestler give the following definition of strong efficient solutions.

DEFINITION 4.3 (see [6]). *An $\bar{x} \in E(f)$ is a strong efficient solution if for any $(x_n)_{n\geq 1} \subset D(f)$ such that $\|f(x_n) - f(\bar{x})\| \to 0$, the sequence $(x_n)_{n\geq 1}$ converges to $\bar{x}$.*

Let $\bar{x} \in E(f)$, $f(\bar{x}) = \eta$, $\epsilon \in \mathcal{K}_0 := \operatorname{int}\mathcal{K} \cup \{0\}$. Let $\pi : \mathcal{K}_0 \rightrightarrows \bar{Z}$ be a set-valued mapping defined as follows:

$$\pi(\epsilon) = f^{-1}(f(\bar{x}) + \epsilon - \mathcal{K}).$$

PROPOSITION 4.4. *Let X and Z be normed spaces and let Z be ordered by a closed cone $\mathcal{K}$, $\operatorname{int}\mathcal{K} \neq \emptyset$. Let $f\colon X \to \bar{Z}$ be a proper function. Let $f(\bar{x}) = \eta$ and $f^{-1}(\eta) = \{\bar{x}\}$. If $\bar{x} \in St^1(f)$, then π is upper Lipschitz at $\epsilon = 0$, i.e., there exist constants $L > 0$ and $t > 0$ such that*

$$f^{-1}(f(\bar{x}) + \epsilon - \mathcal{K}) \subset \bar{x} + L\|\epsilon\|B_X$$

for $\epsilon \in tB_Z \cap \mathcal{K}_0$.

Proof. Suppose that π is not upper Lipschitz at $\epsilon = 0$. For every $n \geq 1$ there exists $\epsilon_n \in \operatorname{int}\mathcal{K}$, $\epsilon_n \to 0$, such that

$$f(x_n) \in f(\bar{x}) + \epsilon_n - \mathcal{K}, \quad \|x_n - \bar{x}\| \geq n\|\epsilon_n\|.$$

Hence, $f(x_n) \in f(\bar{x}) + \frac{1}{n}\|x_n - \bar{x}\|B_Z - \mathcal{K}$ which proves that $\bar{x}$ is not sharp. □

THEOREM 4.5. *If $\bar{x} \in St^1(f)$, then $\bar{x}$ is strong in the sense of Definition 4.3.*

Proof. Let $\bar{x} \in St^1(f)$. Then $\|f(x) + k - f(\bar{x})\| \geq \tau\|x - \bar{x}\|$ for all $x \in D(f)$ and $k \in \mathcal{K}$. In particular $\|f(x) - f(\bar{x})\| \geq \tau\|x - \bar{x}\|$ for all $x \in D(f)$, which proves the assertion. □

In a straightforward way we get the following corollary.

COROLLARY 4.6. *Let X be a Banach space and let Z be a normed space ordered by a closed pointed cone $\mathcal{K}$, $\operatorname{int}\mathcal{K} \neq \emptyset$. Let $f\colon X \to \bar{Z}$ be a quasi-lsc proper function. Let $\varepsilon > 0$ and let $\bar{x} \in D(f)$ be an ε-efficient solution in the direction $k \in \mathcal{K}$.*

For any $k \in \operatorname{int}\mathcal{K}$ there exists $x_k \in X$ such that for any $l \geq 2k$ a point x_k is a strong efficient solution of $f_{(x_k,l)}(x) = f(x) + \frac{\varepsilon}{\lambda}\|x - x_k\|l$.

Proof. Take any $k \in \operatorname{int}\mathcal{K}$ and $l \geq 2k$. By Theorem 4.1 we obtain that $x_k \in St^1(f_{(x_k,l)})$ and therefore by Theorem 4.5 a solution x_k is strong in the sense of Definition 4.3 for any $l \geq 2k$. □

Acknowledgments. We are grateful to the anonymous referees for all the comments and remarks which helped us to improve the results of this paper.

REFERENCES

[1] E. M. BEDNARCZUK, *Weak sharp efficiency and growth condition for vector-valued functions with applications*, Optimization, 53 (2004), pp. 455–474.

[2] F. CAMMAROTO AND A. CHINNI, *A complement to Ekeland's variational principle in Banach spaces*, Bull. Polish Acad. Sci. Math., 44 (1996), pp. 29–33.

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

[4] J. X. FANG, *The variational principle and fixed points theorems in certain topological spaces*, J. Math. Anal. Appl., 208 (1996), pp. 389–412.

[5] C. FINET, *Variational principles in partially ordered Banach spaces*, J. Nonlinear Convex Anal., 2 (2001), pp. 167–174.

[6] C. Finet, L. Quarta, and C. Troestler, *Vector-valued variational principles*, Nonlinear Anal., 52 (2003), pp. 197–218.
[7] P. G. Georgiev, *Parametric Ekeland's variational principle*, Appl. Math. Lett., 14 (2001), pp. 691–696.
[8] P. G. Georgiev, *The strong Ekeland variational principle, the strong drop theorem and applications*, J. Math. Anal. Appl., 131 (1988), pp. 1–21.
[9] A. Göpfert, H. Riahi, C. Tammer, and C. Zalinescu, *Variational Methods in Partially Ordered Spaces*, Springer-Verlag, New York, 2003.
[10] A. Göpfert, C. Tammer, and C. Zalinescu, *On the vectorial Ekeland's variational principle and minimal point theorems in product spaces*, Nonlinear Anal., 39 (2000), pp. 909–922.
[11] G. Isac, *Nuclear cones in product spaces, Pareto efficiency and Ekeland-type variational principle in locally convex spaces*, Optimization, 53 (2004), pp. 253–268.
[12] G. Isac and C. Tammer, *Nuclear and full nuclear cones in product spaces: Pareto efficiency and Ekeland-type variational principle*, Positivity, 9 (2005), pp. 511–539.
[13] W. Oettli and M. Théra, *Equivalents of Ekeland's principle*, Bull. Austral. Math. Soc., 48 (1993), pp. 385–392.
[14] R. Phelps, *Convex Functions, Monotone Operators and Differentiability*, Lecture Notes in Math. 1364, Springer-Verlag, Berlin, Heidelberg, 1989.
[15] R. Phelps, *Support cones in Banach spaces and their applications*, Advances in Math., 13 (1974), pp. 1–19.
[16] C. Tammer, *A generalization of Ekeland's variational principle*, Optimization, 25 (1992), pp. 129–141.
[17] M. Théra, *Etudes des fonctions convexes vectorielles semi-continues*, Thèse de 3ème cycle, Université de Pau, Pau, France, 1978.

SIAM J. OPTIM.
Vol. 18, No. 3, pp. 914–925

DERIVING THE CONTINUITY OF MAXIMUM-ENTROPY BASIS FUNCTIONS VIA VARIATIONAL ANALYSIS*

N. SUKUMAR† AND R. J.-B. WETS‡

Abstract. In this paper, we prove the continuity of maximum-entropy basis functions using variational analysis techniques. The use of information-theoretic variational principles to derive basis functions is a recent development. In this setting, data approximation is viewed as an inductive inference problem, with the basis functions being synonymous with a discrete probability distribution, and the polynomial reproducing conditions acting as the linear constraints. For a set of distinct nodes $\{x^i\}_{i=1}^n$ in $\mathbb{R}^d$, the convex approximation of a function $u(x)$ is $u^h(x) = \sum_{i=1}^n p_i(x)u_i$, where $\{p_i\}_{i=1}^n$ are nonnegative basis functions, and $u^h(x)$ must reproduce affine functions $\sum_{i=1}^n p_i(x) = 1$, $\sum_{i=1}^n p_i(x)x^i = x$. Given these constraints, we compute $p_i(x)$ by minimizing the relative entropy functional (Kullback–Leibler distance), $D(p\|m) = \sum_{i=1}^n p_i(x)\ln\big(p_i(x)/m_i(x)\big)$, where $m_i(x)$ is a known prior weight function distribution. To prove the continuity of the basis functions, we appeal to the theory of epiconvergence.

Key words. maximum entropy, relative entropy, convex approximation, meshfree methods, epiconvergence

AMS subject classifications. 65N30, 65K10, 90C25, 62B10, 26B25

DOI. 10.1137/06066480X

1. Background and formulation.

Consider a set of distinct nodes in $\mathbb{R}^d$ that are located at x^i $(i = 1, 2, \ldots, n)$, with $D = \mathrm{con}(x^1, \ldots, x^n) \subset \mathbb{R}^d$ denoting the convex hull of the nodal set (Figure 1). For a real-valued function $u(x) : D \to \mathbb{R}$, the numerical approximation for $u(x)$ is written as

$$u^h(x) = \sum_{i=1}^{n} p_i(x)u_i, \tag{1}$$

where $p_i(x)$ is the basis function associated with node i, and u_i are coefficients. If $p_i(x)$ is a cardinal basis, $p_i(x^j) = \delta_{ij}$, then $u^h(x^i) = u(x^i) = u_i$.

In the univariate case, Lagrange and spline bases are well known, whereas for multivariate approximation, tensor-product splines, moving least squares (MLS) approximates [17], and radial basis functions [30] are popular. The need for scattered data approximation arises in many fields, for example, curve and surface fitting, computer graphics and geometric modeling, finite elements, and meshfree methods. Over the past decade, meshfree approximation schemes have been adopted in Rayleigh–Ritz (Galerkin) methods for the modeling and simulation of physical phenomena; see [4] for a review of meshfree methods and [28] for a review of meshfree basis functions. For second-order partial differential equations (PDEs), approximates that possess constant and linear precision are sufficient for convergence in a Galerkin method (cf., for

*Received by the editors July 11, 2006; accepted for publication (in revised form) March 26, 2007; published electronically October 4, 2007. This research was supported in part by the National Science Foundation through grants CMMI-0626481 and DMS-0205699.
http://www.siam.org/journals/siopt/18-3/66480.html
†Department of Civil and Environmental Engineering, University of California, Davis, CA 95616 (nsukumar@ucdavis.edu).
‡Department of Mathematics, University of California, Davis, CA 95616 (rjbwets@ucdavis.edu).

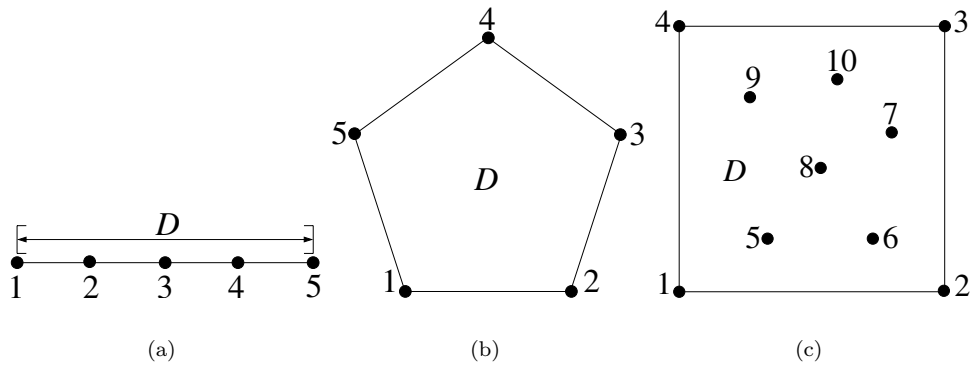


FIG. 1. *Nodal locations* x^i. (a) *One dimension;* (b) *pentagon; and* (c) *scattered nodes within a square.*

example, [25, Chapter 2]):

$$\forall x, \quad \sum_{i=1}^{n} p_i(x) = 1 \quad \text{and} \quad \sum_{i=1}^{n} p_i(x) x^i = x. \tag{2}$$

Furthermore, if the nonnegative restriction is imposed on the basis functions (convex combination), namely,

$$p_i(x) \geq 0 \quad \forall i,\, x, \tag{3}$$

then (1) is a convex approximation scheme [1] with many desirable properties: it satisfies the convex hull property, is not prone to the Runge phenomena, interior nodal basis functions $p_i(x)$ ($x^i \notin \operatorname{bdry} D$) vanish on $\operatorname{bdry} D$, which facilitates the imposition of linear Dirichlet boundary conditions in a Galerkin method, and, in addition, optimal conditioning can be established for nonnegative basis functions [8, 19].

In meshfree Galerkin methods, an approximation of the form in (1) is used, with MLS being the most common choice. A recent development in this direction has been the construction of maximum-entropy approximates [1, 26, 27]; continuity was obtained by Arroyo and Ortiz [1] for the case when the prior distributions are Gaussian. In this paper, we rely on *variational analysis* techniques, in particular on the theory of *epiconvergence*, to establish the continuity of maximum-entropy basis functions for *any* continuous prior distribution.

1.1. Minimum relative entropy principle. In information theory [7], the notion of entropy as a measure of uncertainty or incomplete knowledge was introduced by Shannon [22]. The Shannon entropy of a discrete probability distribution is

$$H(p) = \langle -\ln p \rangle = -\sum_{i=1}^{n} p_i \ln p_i, \tag{4}$$

where $\langle \cdot \rangle$ is the expectation operator, $p_i \equiv p(x^i)$ is the probability of the occurrence of the event x^i, $p \ln p \doteq 0$ if $p = 0$, and the above form of H satisfies the axiomatic requirements of an uncertainty measure; cf., for example, [14, Chapter 1].

Jaynes used the Shannon entropy measure to propose the principle of maximum entropy [11], in which it was shown that maximizing entropy provides the least-biased statistical inference when insufficient information is available. It was later recognized that for H to be invariant under invertible mappings of x, the general form of the entropy should be [12, 15, 23]

$$H(p,m) = -\int p(x)\ln\left(\frac{p(x)}{m(x)}\right)dx \quad \text{or} \quad H(p,m) = -\sum_{i=1}^{n} p_i \ln\left(\frac{p_i}{m_i}\right), \tag{5}$$

where m is a prior distribution that plays the role of a p-estimate. In the literature, the quantity $D(p\|m) = -H(p,m)$ is also referred to as the Kullback–Leibler distance (directed- or I-divergence) [16], and the variational principle is known as the principle of minimum relative entropy [23]. If a uniform prior, $m_i = 1/n$, is used in (5), then the Shannon entropy (modulo a constant) given in (4) is recovered. The nonnegativity of the relative entropy, $D(p\|m) \geq 0$, is readily derived from Jensen's inequality (cf., for example, [7, p. 25]).

Given a set of $\ell + 1$ linear constraints on an unknown probability distribution p and a prior m, which is an estimate for p, the minimum relative entropy principle is a rule for the most consistent (minimum-distance or -discrepancy from the prior m) assignment of the probabilities p_i [12]:

$$\min_{p\in\mathbb{R}^n_+}\left(D(p\|m) = \sum_{i=1}^{n} p_i \ln\left(\frac{p_i}{m_i}\right)\right) \quad \text{so that} \quad \sum_{i=1}^{n} p_i = 1, \tag{6a}$$

$$\sum_{i=1}^{n} p_i g_r(x^i) = \langle g_r(x)\rangle, \quad r = 1, 2, \ldots, \ell, \tag{6b}$$

where $g_r(x)$ and $\langle g_r(x)\rangle$ are known, and $\mathbb{R}^n_+$ is the nonnegative orthant.

The initial emphasis of the principle of maximum entropy was on equilibrium and nonequilibrium statistical mechanics [12], but it is equally applicable to any problem in inductive inference. The interested reader can refer to [13] and [24] for the Bayesian perspective on probability theory and rationale inference. The maximum entropy and minimum relative entropy principles have found applications in many areas of science and engineering—image reconstruction [10], natural language modeling [5], microstructure reconstruction [18], and nonparametric supervised learning [9] are a few examples.

Variational principles, which are used in finite element formulations, conjugate gradient methods, graphical models, dynamic programming, and statistical mechanics, also have strong roots in data approximation. For instance, kriging, thin-plate splines, B-splines, radial basis functions [30], MLS approximates [17], and Delaunay interpolates [20] are based on the extremum of a functional. In the same spirit, we now present the variational formulation to construct entropy approximates, and in so doing demonstrate its potential merits as a basis for the solution of PDEs.

1.2. Variational formulation for entropy approximates. To obtain the maximum-entropy principle, the Shannon entropy functional and a modified entropy functional were used in [26] and [1], respectively. In [27], as a unifying framework and generalization, the relative entropy functional with a prior was used—a uniform prior leads to Jaynes's maximum-entropy principle, and use of a Gaussian (radial basis function) prior, $m_i(x) = \exp(-\beta|x^i - x|^2)$, results in the entropy functional considered in [1]. The prior in the present context is a nodal weight function, and

the variational principle in effect provides a "correction" that minimally modifies the weight functions to form basis functions that also satisfy the linear constraints. Clearly, if $m_i(x)$ a priori satisfies all the constraints, then one obtains $p_i(x) = m_i(x)$ for all i. The flexibility of choosing different prior distributions (for example, radial basis functions, compactly supported weight functions used in MLS, etc.) within the minimum relative entropy formalism would lead to the construction of a wider class of convex approximation schemes. The parallels between the conditions on p_i in (2) and (3) and those on p_i in a maximum-entropy formulation are evident. Unlike univariate Bernstein basis functions (terms in the binomial expansion), where a probabilistic interpretation in relation to the binomial distribution [24, Chapter 5] is natural, here the connection is less transparent. Referring to the nodal sets shown in Figure 1, we note that the basis function value $p_i(x)$ is viewed as the "probability of influence of a node i at x." With a uniform prior, global basis functions are obtained, which do not lead to sparse system matrices in the numerical solution of PDEs. With a compactly supported prior, the basis functions $p_i(x)$ also inherit the support properties of the prior and hence are suitable in the Galerkin solution for PDEs. Entropic regularization with a prior is a novel approach to constructing convex approximation schemes with many desirable properties.

The variational formulation for entropy approximates is as follows: Find $x \mapsto p(x) : \mathbb{R}^d \to \mathbb{R}^n_+$ as the solution of the constrained convex optimization problem

$$\min_{p \in \mathbb{R}^n_+} f(x;p), \quad f(x;p) = \sum_{i=1}^{n} p_i(x) \ln\left(\frac{p_i(x)}{m_i(x)}\right), \tag{7a}$$

subject to the constraint set from (2) and (3):

$$\kappa(x) = \left\{ p \in \mathbb{R}^n_+ \,\middle|\, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i x^i = x \right\}, \tag{7b}$$

where $m_i(x)$ is a prior estimate, and the constraints form an underdetermined linear system. By introducing the Lagrange multipliers, one can write the solution of the variational problem as

$$p_i(x) = \frac{Z_i(x)}{Z(x)}, \quad Z_i(x) = m_i(x) \exp(-x^i \cdot \lambda),$$

where $\lambda \in \mathbb{R}^d$, and $Z(x) = \sum_j Z_j(x)$ is known as the partition function in statistical mechanics. The $p_i(x)$ in the preceding equation must satisfy the d linear constraints in (7b). This yields d nonlinear equations. On using shifted nodal coordinates $\tilde{x}^i = x^i - x$ and considering the dual formulation, we can write the solution for the Lagrange multipliers as (cf., for example, [21, Exercise 11.12] and [6, p. 222])

$$\lambda = \arg\min \ln Z(\lambda^t),$$

where Z is appropriately redefined. Convex optimization algorithms (gradient descent, Newton's method, etc.) are suitable for computing these basis functions. Numerical experimentation suggests that such basis functions may very well be continuous on D [1, 26], and this will be confirmed here by variational analysis techniques.

2. Continuity of the basis functions. One can always represent an optimization problem, involving constraints or not, as one of minimizing an extended real-valued function. In the case of a constrained-minimization problem, simply redefine the effective objective as taking on the value ∞ outside the feasible region, with the set determined by the constraints. In this framework, the canonical problem can be formulated as one of minimizing on all of $\mathbb{R}^n$ an extended real-valued function $f\colon \mathbb{R}^n \to \overline{\mathbb{R}}$. Approximation issues can consequently be studied in terms of the convergence of such functions. This has led to the notion of *epiconvergence* (cf. [2, 3] and [21, Chapter 7]; the latter will serve here as our basic reference). We provide a very brief survey and some relevant refinements of this theory.

Thus, at a conceptual level, it is convenient to think of optimization problems as elements of

$$\operatorname{fcns}(\mathbb{R}^n) = \{f\colon \mathbb{R}^n \to \overline{\mathbb{R}}\},$$

the set of extended real-valued functions that are defined on *all* of $\mathbb{R}^n$, even allowing for the possibility that they are nowhere finite valued; definitions, properties, limits, etc., usually do not refer specifically to the domain on which they are finite. The *effective domain* of f is $\operatorname{dom} f = \{x \in \mathbb{R}^n \mid f(x) < \infty\}$. The *epigraph* of a function f is the set of all points in $\mathbb{R}^{n+1}$ that lie on or above the graph of f, $\operatorname{epi} f = \{(x,\alpha) \in \mathbb{R}^{n+1} \mid \alpha \geq f(x)\}$. A function f is *lsc* (*lower semicontinuous*) if and only if its epigraph is closed as a subset of $\mathbb{R}^{n+1}$, i.e., $\operatorname{epi} f = \operatorname{cl}(\operatorname{epi} f)$ with cl denoting closure [21, Theorem 1.6]. The lsc-regularization of f is $\operatorname{cl} f$ defined by the identity $\operatorname{epi} \operatorname{cl} f = \operatorname{cl} \operatorname{epi} f$.

DEFINITION 2.1 (epiconvergence and tight epiconvergence). *Let* $\{f, f^\nu, \nu \in \mathbb{N}\}$ *be a collection of functions in* $\operatorname{fcns}(\mathbb{R}^n)$. *Then,* $f^\nu \xrightarrow{e} f$ *if and only if the following conditions are satisfied:*

(a) *For all* $x^\nu \to x$, $\operatorname{liminf}_\nu f^\nu(x^\nu) \geq f(x)$.

(b) *For all* x, $\exists\, x^\nu \to x$ *such that* $\operatorname{limsup}_\nu f^\nu(x^\nu) \leq f(x)$.

The sequence epiconverges tightly *to* f *if, in addition, for all* $\epsilon > 0$, *there exist a compact set* B_ϵ *and an index* ν_ϵ *such that*

$$\forall\, \nu \geq \nu_\epsilon : \quad \inf\nolimits_{B_\epsilon} f^\nu \leq \inf f^\nu + \epsilon.$$

Note that functions can be "epiclose" while "pointwise-far" (measured, for example, in term of the ℓ^∞-norm); e.g., consider the two step-functions $f(x) = 0$ if $x < 0$, $f(x) = 1$ when $x \geq 0$, and $g(x) = f(x - \epsilon)$ with $\epsilon > 0$ arbitrarily small.

The name "epiconvergence" is attached to this convergence notion because it coincides [21, Proposition 7.2] with the *set-convergence*, in the Painlevé–Kuratowski sense [21, section 4.B] of the epigraphs. It is known that (i) whenever C is a limit-set, it is *closed* [21, Proposition 4.4]; (ii) $C = \emptyset$ if and only if the sequence C^ν eventually "escapes" from any bounded set [21, Corollary 4.11]; and (iii) if the sequence $C^\nu \to C$ consists of convex sets, then also C is convex [21, Proposition 4.15]. This means that when $f^\nu \xrightarrow{e} f$, (i) f is lsc; (ii) $f \equiv \infty (\operatorname{dom} f = \emptyset)$ if and only if given any $\kappa > 0$, $f^\nu \geq \kappa$ for ν large enough; and (iii) the epilimit of convex functions is convex, if it exists.

THEOREM 2.2 (convergence of the minimizers and infimums). *Let* $f^\nu \xrightarrow{e} f$, *all in* $\operatorname{fcns}(\mathbb{R}^n)$, *with* $\inf f$ *finite. If* $f^\nu \xrightarrow{e} f$, $x^k \in \operatorname{argmin} f^{\nu_k}$ *for some subsequence* $\{\nu_k\}_{k\in\mathbb{N}}$ *and* $x^k \to \bar{x}$, *then* $\bar{x} \in \operatorname{argmin} f$ *and* $\min f^{\nu_k} \to \min f$.[1]

[1] One writes min when the infimum is actually attained.

If $\operatorname{argmin} f$ *is a singleton, then every convergent subsequence of minimizers converges to* $\operatorname{argmin} f$.

They epiconverge tightly if and only if $\inf f^\nu \to \inf f$.

Proof. The first two assertions follow from [21, Proposition 7.30, Theorem 7.33], and one can deduce the last one from [21, Theorem 7.31]. ∎

Let us conclude this review by a compilation of the facts that are going to be of immediate relevance to the problem at hand.

COROLLARY 2.3 (epiconvergence under strict convexity). *Suppose* $\{f^\nu : \mathbb{R}^n \to (-\infty, \infty]\}_{\nu\in\mathbb{N}}$ *is a collection of convex functions such that*

(a) *for all* ν, $\operatorname{dom} f^\nu \subset B$, *where* B *and each* $\operatorname{dom} f^\nu$ *are compact;*

(b) *the functions* f^ν *are finite valued, lsc, and strictly convex on* $\operatorname{dom} f^\nu$. *Then, for all* ν, $\emptyset \neq \operatorname{argmin} f^\nu$ *is a singleton.*

Moreover, if $f^\nu \xrightarrow{e} f$ *and* $\operatorname{argmin} f$ *is also a singleton, then* $\operatorname{argmin} f^\nu \to \operatorname{argmin} f$.

Proof. In view of (a) and (b), for each ν the minimization of f^ν is equivalent to minimizing a finite-valued, lsc, strictly convex function on a compact set, and such a problem always has a unique solution. Moreover, because for all ν, $\operatorname{dom} f^\nu$ is a (compact) subset of the compact set B, $f^\nu \xrightarrow{e} f$ implies that they epiconverge tightly. The convergence of $\operatorname{argmin} f^\nu \to \operatorname{argmin} f$ follows from combining the two last assertions of Theorem 2.2. ∎

Our task now is to show that the continuity of the basis functions can be derived as a consequence of this corollary. We begin with the strict convexity of the criterion function. The Kullback–Leibler criterion is a separable function, i.e.,

$$k(x;p) = \sum_{i=1}^{n} k_i(x;p_i), \quad \text{where } k_i(x;p_i) = p_i \ln(p_i/m_i(x)),$$

and its properties can be directly derived from those of the one-dimensional functions $k_i(x;\cdot) : \mathbb{R}_+ \to [0,\infty]$.

- When $m_i(x) > 0$, $k_i(x,\cdot)$ is finite valued, continuous, and strictly convex on $\mathbb{R}_+$; recall that $0\ln(0) = 0$. Indeed, the second derivative on $(0,\infty)$ is $1/p_i > 0$, which implies strict convexity [21, Theorem 2.13(c)]. The quantity $p_i \ln(p_i/m_i(x))$ is strictly increasing and converges to 0 as $p_i \searrow 0$, yielding both strict convexity and continuity on $\mathbb{R}_+$.
- When $m_i(x) = 0$, $k_i(x;p_i) = \infty$ unless $p_i = 0$ and then $k_i(x;0) = 0$.

It is conceivable, but certainly not reasonable, that the (continuous) weight functions $\{m_i : \mathbb{R}^n \to \mathbb{R}_+,\ i = 1,\dots,n\}$ have been chosen so that for some $x \in D$, $m_i(x) = 0$ for all $i = 1,\dots,n$. In such a situation, in the process of minimizing the Kullback–Leibler criterion, we would be led to choosing $p = 0$ and, of course, this would make it impossible to satisfy the constraint $\sum_{i=1}^n p_i = 1$; i.e., the problem, so formulated, would be infeasible! This brings us to the following assumption, in which we let

- $\text{s-supp}\, m_i = \{x \in \mathbb{R}^d \,|\, m_i(x) > 0\}$ denote the *strict support* of m_i, and
- $\operatorname{supp} m_i = \operatorname{cl}(\text{s-supp}\, m_i)$ the *support* of m_i.

ASSUMPTION 2.1 (well-posed assumption). *For each* $i = 1,\dots,n$, *the function* $m_i : \mathbb{R}^n \to \mathbb{R}_+$ *is continuous such that* $\text{s-supp}\, m_i$, *and consequently also* $\operatorname{supp} m_i$, *is nonempty.*[2] *Moreover, with* $I_{=0} = \{i \,|\, m_i(x) = 0\}$ *and* $I_{>0} = \{i \,|\, m_i(x) > 0\}$,

$$\forall\, x \in D: \quad x \in \operatorname{con}(x^i \,|\, i \in I_{>0}).$$

[2]Note that the continuity of m_i implies that $\text{s-supp}\, m_i$ is an open subset of $\mathbb{R}^d$, and thus so is $\bigcup_{i=1}^n \text{s-supp}\, m_i$.

This assumption requires that every $x \in D$ be obtained as a convex combination of some subcollection of the nodal locations x^i that are associated with weight functions m_i that have $m_i(x) > 0$. In particular, this implies that $\kappa(x)$ is never empty, or equivalently, that the constraints (7b) are certainly satisfied whenever $x \in D$.

PROPOSITION 2.4 (the Kullback–Leibler criterion). *Under the well-posed Assumption* 2.1*, for all $x \in D$, the Kullback–Leibler criterion $p \mapsto k(x;p) = \sum_{i=1}^n p_i \ln (p_i/m_i(x))$ is a strictly convex, lsc function on $\mathbb{R}^n_+$, taking into account the identity $0 \ln(0) = 0$.*

Proof. Convexity is well known; see [7, p. 30], [21, Exercise 3.51], for example. Again, with $I_{=0} = \{i \mid m_i(x) = 0\}$ and $I_{>0} = \{i \mid m_i(x) > 0\}$,

$$k(x;p) = \sum_{i \in I_{=0}} k_i(x;p_i) + \sum_{i \in I_{>0}} k_i(x;p_i),$$

$\operatorname{dom} k(x;\cdot) = \prod_{I_{=0}} \{0\} \times \prod_{I_{>0}} \mathbb{R}_+$, and $I_{>0}$ nonempty by Assumption 2.1. From our analysis of the functions $k_i(x;\cdot)$, it follows that $k(x;\cdot)$ is strictly convex, continuous on its effective domain $\operatorname{dom} k(x;\cdot)$. ∎

The tools are now at hand to derive our main result.

THEOREM 2.5 (continuity of the basis functions). *For $x \in D$, as in the formulation of maximum entropy* (7)*, let*

$$f(x;p)) = \begin{cases} \sum_{i=1}^n p_i \ln(p_i/m_i(x)) & \text{if } p \in \kappa(x), \\ \infty & \text{otherwise,} \end{cases}$$

where

$$\kappa(x) = \left\{ p \in \mathbb{R}^n_+ \,\middle|\, \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i x^i = x \right\},$$

and where

$$p(x) = (p_1(x), \ldots, p_n(x)) = \operatorname{argmin} f(x;\cdot).$$

Under the well-posed Assumption 2.1*, when $x^\nu \to \bar{x}$ with $x^\nu \in D$, $\kappa(\bar{x})$ is nonempty and*

$$f(x^\nu;\cdot) \xrightarrow{e} f(\bar{x};\cdot) \quad \text{and} \quad p(x^\nu) \to p(\bar{x}).$$

In other words, the basis functions $p(\cdot)$ are continuous on D.

Proof. Since for all $x \in D$, $\kappa(x)$ is a compact, nonempty subset of the unit simplex $\Delta = \{p \in \mathbb{R}^n_+ \mid \sum_{i=1}^n p_i = 1\}$, it follows that for all $x \in D$, $\operatorname{dom} f(x;\cdot) \subset \Delta$ and, consequently, condition (a) of Corollary 2.3 is trivially satisfied. The rest of the proof is concerned with condition (b) and the epiconvergence of the sequence $f(x^\nu;\cdot)$ to $f(\bar{x};\cdot)$ when $x^\nu \to \bar{x}$.

The functions $f(x^\nu;\cdot)$ and $f(\bar{x};\cdot)$ can be written as $k(x^\nu;\cdot) + \iota_{\kappa(x^\nu)}$ and $k(\bar{x};\cdot) + \iota_{\kappa(\bar{x})}$, where $k(x;p)$ is the Kullback–Leibler criterion defined on $\mathbb{R}^n_+$ and ι_C is the indicator function of the set $C \subset \mathbb{R}^n$ with $\iota_C = 0$ on C; otherwise, $\iota_C = \infty$ on $\mathbb{R}^n \setminus C$.

The epiconvergence of $f(x^\nu;\cdot)$ to $f(\bar{x};\cdot)$ follows from [21, Theorem 7.46(b)], which asserts that the sum of two sequences of functions epiconverge to the sum of their limits if one sequence epiconvergences and the other converges continuously.

To obtain the epiconvergence of the indicator functions, or equivalently [21, Proposition 7.4(f)] the set convergence of the sets $\kappa(x^\nu) \to \kappa(\bar{x})$ with $\kappa(\bar{x}) \neq \emptyset$, we exploit the fact that these are polyhedral sets and that, on the bounded polyhedral set $D = \mathrm{con}(x^1, \ldots, x^n) \subset \mathbb{R}^d$, the mapping $x \mapsto \kappa(x)$ is Lipschitz continuous with respect to the Pompeiu–Hausdorff distance $d\!\!\!l_\infty$, i.e.,

$$\forall\, x, x' \in D\colon \quad d\!\!\!l_\infty(\kappa(x), \kappa(x')) \leq M|x - x'|$$

for some constant $M > 0$; here $|\cdot|$ denotes the Euclidean norm; cf. [29, Theorem 1]; see also [21, Example 9.35]. Of course, this means that κ is continuous on D and, in particular, for any $x^\nu \to \bar{x}$ in D, given any sequence $p^\nu \in \kappa(x^\nu) \to \bar{p}$, then $\bar{p} \in \kappa(\bar{x})$.

Thus, to assert continuous convergence of the functions $k(x^\nu;\cdot)$ to $k(\bar{x})$, one needs to show that $k(x^\nu;p^\nu) \to k(\bar{x};\bar{p})$ for such pairs (x^ν, p^ν). Let $I_{=0} = \{i \,|\, m_i(\bar{x}) = 0\}$ and $I_{>0} = \{i \,|\, m_i(\bar{x}) > 0\}$. By Assumption 2.1, $\kappa(\bar{x}) \bigcap (\bigcup_{I_{>0}} \text{s-supp}\, m_i) \neq \emptyset$. Furthermore, the open set $\bigcup_{I_{>0}} \text{s-supp}\, m_i$ not only includes $\bar{x}$ but also x^ν for all ν large enough. Thus, for all $i \in I_{>0}$, $p_i^\nu \ln(p_i^\nu)/m_i(x^\nu) \to \bar{p}_i \ln(\bar{p}_i)/m_i(\bar{x})$. When $i \in I_{=0}$, again for ν large enough, $p_i^\nu = 0 = \bar{p}_i$; otherwise the corresponding vectors p^ν and $\bar{p}$ would not belong to $\mathrm{dom}\, k(x^\nu;\cdot)$ or $\mathrm{dom}\, k(\bar{x};\cdot)$. Hence, $k(x^\nu;p^\nu) \to k(\bar{x};\bar{p})$. So, $f(x^\nu;\cdot) \to f(\bar{x},\cdot)$.

There only remains to observe that, for ν large enough, $\mathrm{argmin}\, f(x^\nu;\cdot)$ is unique, i.e., for $i \notin I_{>0}$, $p_i^\nu(x^\nu) = 0$, whereas for $i \in I_{>0}$, $p_i^\nu(x^\nu) = \mathrm{argmin}_{p_i \geq 0}\, p_i \ln(p_i/m_i(x^\nu))$; the strict convexity guarantees that argmin is a singleton. Since the same holds for $\bar{x}$, we are in the framework of Corollary 2.3, and thus $p(x^\nu) = \mathrm{argmin}\, f(x^\nu;\cdot) \to \mathrm{argmin}\, f(\bar{x};\cdot) = p(\bar{x})$. ∎

3. Numerical experiments. To illustrate Theorem 2.5, we present basis function plots to confirm the continuity of maximum-entropy basis functions. First, one-dimensional basis function plots are considered, and then two-dimensional basis function plots are presented.

To demonstrate a simple closed-form computation, consider one-dimensional approximation in $D = [0, 1]$ with three nodes located at $x_1 = 0$, $x_2 = 1/2$, and $x_3 = 1$. On using (7), the solution for $p_i(x)$ is obtained by solving a quadratic equation:

$$p_1(x) = \frac{1}{Z}, \quad p_2(x) = \frac{\eta}{Z}, \quad p_3(x) = \frac{\eta^2}{Z}, \quad \eta \equiv \eta(x) = \frac{2x - 1 + \sqrt{12x(1-x)+1}}{4(1-x)},$$

where $Z = 1 + \eta + \eta^2$. These basis functions are presented in Figure 2(a). For four equispaced nodes in $[0, 1]$, a cubic equation must be solved. In general, a numerical method is required to compute these basis functions; in our computations, we use a one-dimensional MATLAB implementation, whereas in two dimensions, a gradient descent algorithm [26, p. 2165] is adopted. Figure 2 depicts basis function plots on a uniform grid consisting of three nodes and five nodes (nodal locations are shown in Figure 1(a)). The plots are presented for a Gaussian prior distribution, $m_i(x) = \exp(-\beta(|x^i - x|^2)$, with varying β. The value $\beta = 0$ corresponds to a uniform prior, and for large β (theoretically when $\beta \to \infty$), the entropy basis functions tend to the finite element Delaunay interpolant [1]. From Figures 2(a) and 2(d), we observe that nodal interpolation is realized on the boundary but not at the interior nodes. However, as β is increased, the support of the basis functions shrinks and the basis functions become closer to being an interpolant at the interior nodes. For $\beta = 100$, the entropy basis functions are proximal to piecewise linear finite element basis functions (Figures 2(c) and 2(f)). The plots in Figure 2 evince the continuity of the basis functions, which provides numerical evidence in support of the theoretical proof in Theorem 2.5.

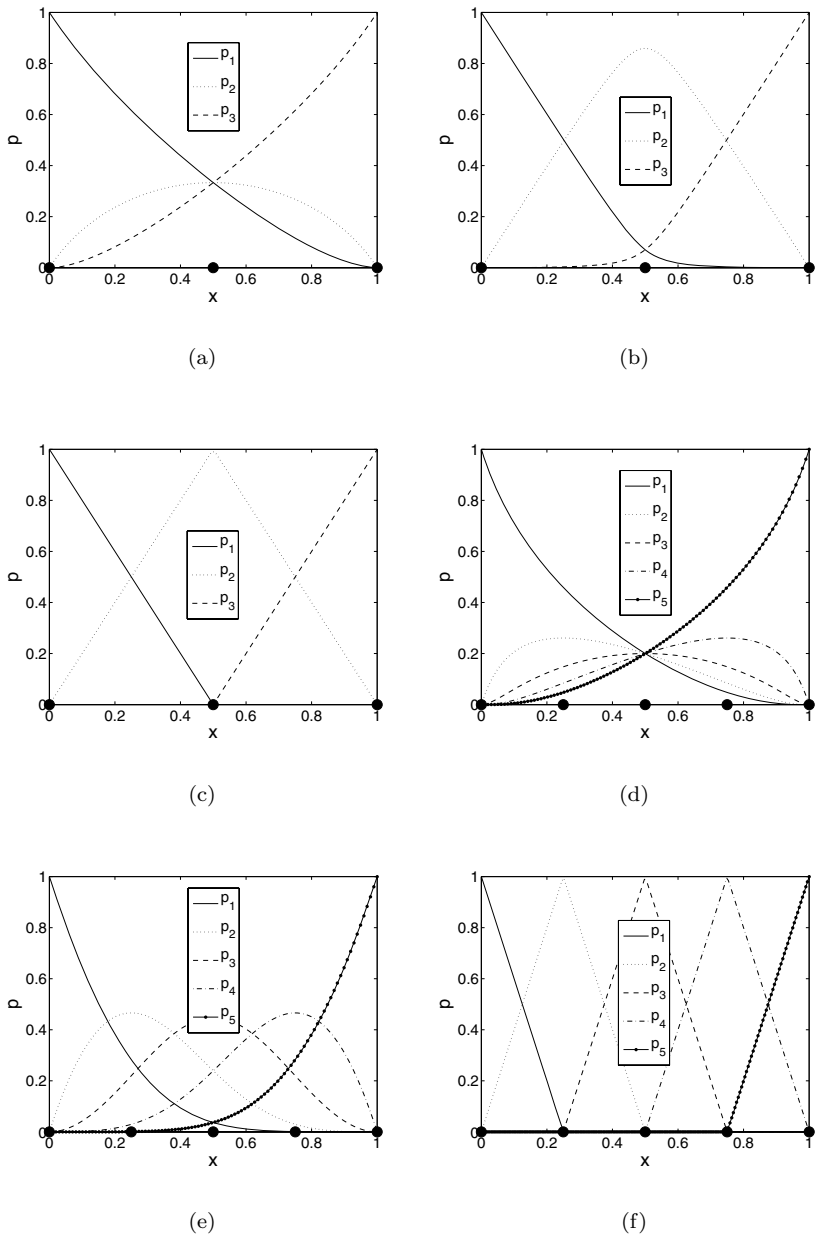
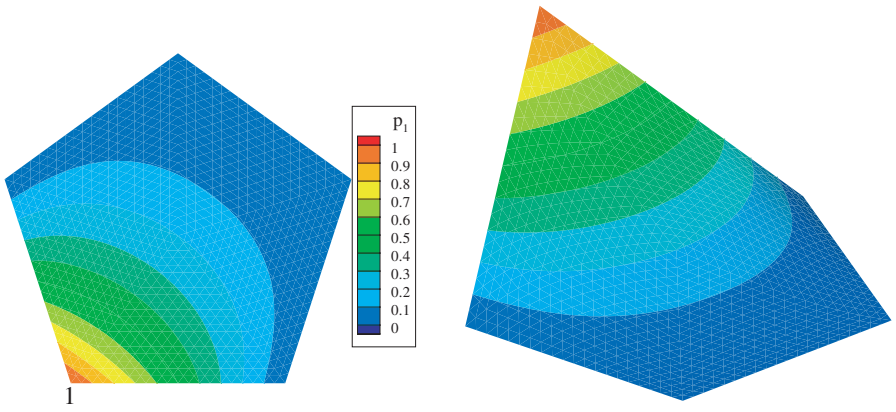
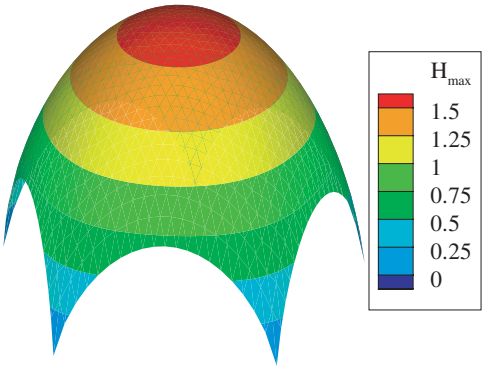


FIG. 2. *Entropy basis functions with a Gaussian prior.* (a)–(c) *show* $n = 3$ *and* $\beta = 0,\ 10,\ 100$*; and* (d)–(f) *show* $n = 5$ *and* $\beta = 0,\ 10,\ 100$. *The nodal locations along the* x*-axis are depicted by filled circles.*

In Figure 3(a), a contour plot of $p_1(x)$ for node 1 in a regular pentagon (see Figure 1(b) for the nodal locations) is shown, whereas in Figure 3(b), the three-dimensional plot is illustrated. The variation of the maximum entropy within the pentagon is depicted in Figure 3(c), with the maximum value of $\ln 5$ being attained at the centroid of the pentagon. The basis function $p_1(x)$ satisfies the cardinal property,



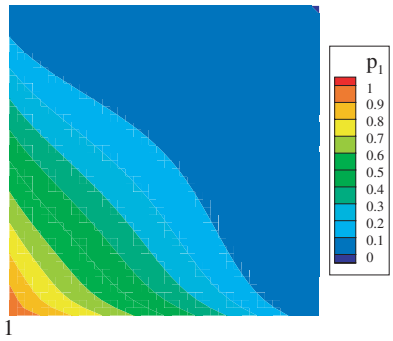
(a) (b)



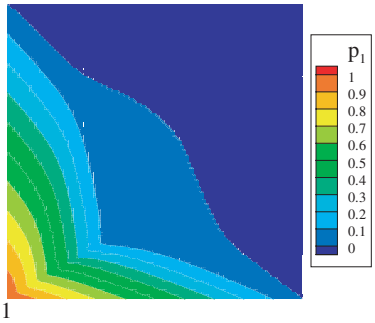
(c)

FIG. 3. *Entropy basis function* $p_1(x)$ *and variation of maximum entropy within a regular pentagon.* (a) *Contour plot;* (b) *three-dimensional plot; and* (c) H_{max}.

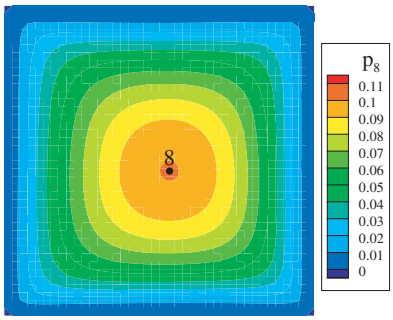
$p_i(x^j) = \delta_{ij}$, which is also met by all n nodal basis functions in a convex polygon [26]. Next, we consider the grid shown in Figure 1(d), where $D = [0,1]^2$. The basis functions for nodes 1 and 8 are plotted using a uniform prior, a Gaussian prior with $\beta = 20$, and a compactly supported C^2 quartic radial basis function as a prior. The quartic prior is given by $m_i(r) = 1 - 6r^2 + 8r^3 - 3r^4$ if $r = |x^i - x| \leq 1$, and zero otherwise. The contour plots are illustrated in Figure 4, and once again we observe that the basis functions are continuous in D. Furthermore, the interior basis functions (for example, $p_8(x)$) vanish on bdry D, which enables the direct imposition of Dirichlet boundary conditions in Galerkin methods [1]. The one- and two-dimensional basis function plots provide numerical proof in support of Theorem 2.5, thereby establishing the continuity of $p_i(x)$ for $x \in D$.



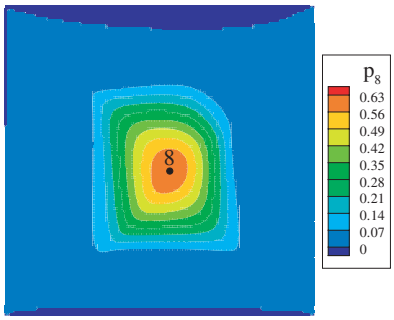
(a)



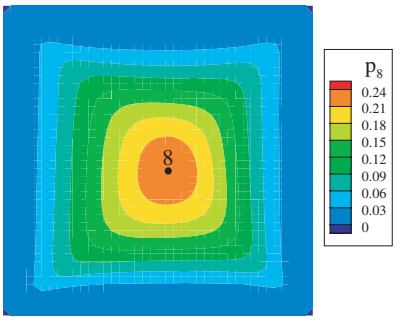
(b)



(c)



(d)



(e)

FIG. 4. *Two-dimensional entropy basis functions within a unit square.* (a) *and* (b) *show* $p_1(x)$ *with a uniform prior and a Gaussian prior* ($\beta = 20$)*;* (c) *and* (d) *show* $p_8(x)$ *with a uniform prior and a Gaussian prior* ($\beta = 20$)*; and* (e) *shows* $p_8(x)$ *with a compactly supported* C^2 *radial basis function.*

REFERENCES

[1] M. Arroyo and M. Ortiz, *Local maximum-entropy approximation schemes: A seamless bridge between finite elements and meshfree methods*, Int. J. Numer. Methods Engrg., 65 (2006), pp. 2167–2202.

[2] H. Attouch, *Variational Convergence for Functions and Operators*, Appl. Math. Ser., Pitman, Boston, 1984.

[3] J.-P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhäuser Boston, Boston, MA, 1990.

[4] T. Belytschko, Y. Krongauz, D. Organ, M. Fleming, and P. Krysl, *Meshless methods: An overview and recent developments*, Comput. Methods Appl. Mech. Engrg., 139 (1996), pp. 3–47.

[5] A. L. Berger, S. A. Della Pietra, and V. J. Della Pietra, *A maximum entropy approach to natural language processing*, Comput. Linguist., 22 (1996), pp. 39–71.

[6] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.

[8] R. T. Farouki and T. N. T. Goodman, *On the optimal stability of the Bernstein basis*, Math. Comp., 65 (1996), pp. 1553–1566.

[9] M. R. Gupta, *An Information Theory Approach to Supervised Learning*, Ph.D. thesis, Department of Electrical Engineering, Stanford University, Palo Alto, CA, 2003.

[10] J. Skilling and R. K. Bryan, *Maximum entropy image reconstruction: General algorithm*, Monthly Notices Roy. Astronom. Soc., 211 (1984), pp. 111–118.

[11] E. T. Jaynes, *Information theory and statistical mechanics*, Phys. Rev., 106 (1957), pp. 620–630.

[12] E. T. Jaynes, *Information Theory and Statistical Mechanics*, in Statistical Physics: The 1962 Brandeis Lectures, K. Ford, ed., W. A. Benjamin, New York, 1963, pp. 181–218.

[13] E. T. Jaynes, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK, 2003.

[14] A. Khinchin, *Mathematical Foundations of Information Theory*, Dover, New York, 1957.

[15] S. Kullback, *Information Theory and Statistics*, Wiley, New York, 1959.

[16] S. Kullback and R. A. Leibler, *On information and sufficiency*, Ann. Math. Statist., 22 (1951), pp. 79–86.

[17] P. Lancaster and K. Salkauskas, *Surfaces generated by moving least squares methods*, Math. Comp., 37 (1981), pp. 141–158.

[18] R. W. Minich, C. A. Schuh, and M. Kumar, *Role of topological constraints on the statistical properties of grain boundary networks*, Phys. Rev. B, 66 (2003), 052101.

[19] J. M. Peña, *B-splines and optimal stability*, Math. Comp., 66 (1997), pp. 1555–1560.

[20] V. T. Rajan, *Optimality by the Delaunay triangulation in R^d*, Discrete Comput. Geom., 12 (1994), pp. 189–202.

[21] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, Heidelberg, 2004.

[22] C. E. Shannon, *A mathematical theory of communication*, Bell Systems Tech. J., 27 (1948), pp. 379–423.

[23] J. E. Shore and R. W. Johnson, *Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy*, IEEE Trans. Inform. Theory, 26 (1980), pp. 26–36.

[24] D. S. Sivia, *Data Analysis: A Bayesian Tutorial*, Oxford University Press, Oxford, UK, 1996.

[25] G. Strang and G. Fix, *An Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs, NJ, 1973.

[26] N. Sukumar, *Construction of polygonal interpolants: A maximum entropy approach*, Int. J. Numer. Methods Engrg., 61 (2004), pp. 2159–2181.

[27] N. Sukumar, *Maximum entropy approximation*, AIP Conf. Proc., 803 (2005), pp. 337–344.

[28] N. Sukumar and R. W. Wright, *Overview and construction of meshfree basis functions: From moving least squares to entropy approximants*, Int. J. Numer. Methods Engrg., 70 (2007), pp. 181–205.

[29] D. Walkup and R. J.-B. Wets, *A Lipschitzian characterization of convex polyhedra*, Proc. Amer. Math. Soc., 23 (1969), pp. 167–173.

[30] H. Wendland, *Scattered Data Approximation*, Cambridge University Press, Cambridge, UK, 2005.

SIAM J. OPTIM.
Vol. 18, No. 3, pp. 926–945

STABILITY ANALYSIS FOR NONLINEAR OPTIMAL CONTROL PROBLEMS SUBJECT TO STATE CONSTRAINTS*

K. MALANOWSKI†

Abstract. There is considered a family of nonlinear optimal control problems depending on a functional parameter. The problems are subject to state constraints. Conditions are derived under which the solutions and Lagrange multipliers are locally Lipschitz continuous functions of the parameter. Those conditions consist of constraint qualifications and weakened coercivity conditions.

Key words. optimal control, nonlinear ODEs, state constraints, parametric problems, second order sufficient conditions, Lipschitz stability of the solutions

AMS subject classifications. 49K40, 49K30, 49K15

DOI. 10.1137/060661247

1. Introduction.

In stability analysis for optimal control problems, conditions are investigated under which the solutions and Lagrange multipliers are locally Lipschitz continuous functions of the parameter. It is know that two types of assumptions are crucial in that analysis: constraint qualifications and a second order optimality condition (coercivity of the Hessian of Lagrangian). For control-constrained problems, a full characterization of stability properties was obtained in [4] (see also [9]) under the provision that the parameters are functions of time and the dependence of data on the parameter is strong. The situation was different for state-constrained problems. In papers [8, 3], devoted to stability analysis for that class of problems, a strong second order optimality condition (coercivity condition) was assumed, where the active state constraints were ignored. In the recent paper [11] of the author, an example was constructed in which the Lipschitz stability property was satisfied, but the strong second order condition of [8, 3] was violated. It shows that this condition is too strong. Note that, in the above mentioned example, the parameter is a function of time. In [11] a new second order condition was introduced, which was weakened by taking into account the strongly active state constraints, i.e., those active constraints for which the pointwise strict complementarity condition was satisfied with the margin $\alpha > 0$. It was shown in [11] and [12] that, under this weakened condition, the solutions and Lagrange multipliers for linear-quadratic optimal control problems remain stable under small perturbations. The crucial point in the proof of this result was the analysis of stability of the coercivity condition in a neighborhood of the reference point.

In the present paper, we still further weaken the second order condition used in [11] and we extend the stability results to nonlinear problems. Thus, our assumptions are essentially weaker than those in [8] and [3]. This weakening turns out to be crucial. As it will be shown in a forthcoming paper of the author, the new conditions are not only sufficient, but also necessary for Lipschitz stability and directional differentiability

*Received by the editors May 29, 2006; accepted for publication (in revised form) March 27, 2007; published electronically October 4, 2007. This research was supported by the Polish Ministry of Scientific Research and Information Technology grant 3 T11C 051 28.
http://www.siam.org/journals/siopt/18-3/66124.html
†Systems Research Institute, Polish Academy of Sciences, ul.Newelska 6, 01–447 Warsaw, Poland (kmalan@ibspan.waw.pl).

of the solutions and Lagrange multipliers for a class of parametric optimal control problems, subject to the first order state constraints.

In the stability analysis for nonlinear optimal control problems, we follow [8] and [3] and use the implicit function theorem for strongly regular generalized equations [16, 2]. The crucial point in this approach is to show that the stationary points of the accessory problems are Lipschitz stable under sufficiently regular small perturbations. We use here the approach similar to that developed for linear-quadratic problems in [11].

The organization of this paper is the following. In section 2, some stability results for cone-constrained optimization problems are recalled. In section 3, the considered optimal control problem is formulated and the needed assumptions are introduced. Some basic results concerning state-constrained optimal control problems are presented. In particular, they include a coercivity condition, which is weakened by taking into account the strongly active state constraints. Following the idea developed in the paper [1], it is shown in the appendix that this coercivity condition constitutes a second order sufficient optimality condition. In section 4 the basic auxiliary lemmas and stability results are formulated. They are proved in section 5.

We use the following notations: Capital letters X, Y, V, etc., with superscripts denote Banach or Hilbert spaces. The norms are denoted by $\|\cdot\|$ with a subscript referring to the space. $\mathcal{B}_\rho^X(x_0) := \{x \in X \mid \|x - x_0\|_X < \rho\}$ is the open ball in X of radius ρ, centered at x_0. Asterisks denote dual spaces as well as adjoint operators. For $f : X \times Y \to Z$, $D_x f(x,y), D_y f(x,y), D^2_{x,y} f(x,y)$, etc., denote the Fréchet derivatives in the respective variables.

$\mathbb{R}^n$ is the n-dimensional Euclidean space, with the inner product denoted by $\langle x, y\rangle$ and the norm $|x| = \langle x, x\rangle^{\frac{1}{2}}$. Transposition is denoted by $*$.

$L^p(0,1;\mathbb{R}^n)$ is the Banach space of measurable functions $f : [0,1] \to \mathbb{R}^n$, supplied with the norm

$$\|f\|_p = \begin{cases} \left[\int_0^1 |f(t)|^p dt\right]^{1/p} & \text{for } p \in [1,\infty), \\ \operatorname{ess\,sup}_{t\in[0,1]} |f(t)| & \text{for } p = \infty. \end{cases}$$

$W^{1,p}(0,1;\mathbb{R}^n)$ denotes the Sobolev space of functions f absolutely continuous on $[0,1]$ with the norm

$$\|f\|_{1,p} = \begin{cases} [|f(0)|^p + \|\dot f\|_p^p]^{1/p} & \text{for } p \in [1,\infty), \\ \max\{|f(0)|, \|\dot f\|_\infty\} & \text{for } p = \infty. \end{cases}$$

We will need the subspace $W_0^{1,p}(0,1;\mathbb{R}^n) := \{f \in W^{1,p}(0,1;\mathbb{R}^n) \mid f(0) = 0\}$.

The inner product and the norm in $L^2(0,1;\mathbb{R}^n)$ are denoted by $(\cdot,\cdot)$ and $\|x\|_2 = (x,x)^{1/2}$, respectively. Similarly, for $W^{1,2}(0,1;\mathbb{R}^n)$ we denote $(\cdot,\cdot)_{1,2}$ and $\|x\|_{1,2} = (x,x)_{1,2}^{1/2}$. c, k, l, and ℓ are generic constants, not necessarily the same in different places.

2. Stability results for cone-constrained problems. In this section we recall some results for parametric cone-constrained optimization problems, which will be used in stability analysis for optimal control problems.

Let $H \subset Z, X$, and Y be Banach spaces of parameters, arguments, and constraints, respectively, where the inclusion $H \subset Z$ is dense and continuous. In the space Y there is given a closed convex cone $\mathcal{K}$, which induces a partial order in Y.

Further, $\mathcal{G}: X \times Z \to \mathbb{R}$ and $\phi: X \times Z \to Y$ are given functions. Consider the family of the following optimization problems depending on the parameter $h \in H$:

$$(\mathrm{P})_h \qquad \min_{\xi \in X} \mathcal{G}(\xi, h) \text{ subject to } \phi(\xi, h) \in \mathcal{K}.$$

Assume the following:

(B1) For each $h \in H$ the functions $\mathcal{G}(\cdot, h)$ and $\phi(\cdot, h)$ are Fréchet differentiable on X.

Let us introduce the following Lagrangian for $(\mathrm{P})_h$:

$$L: X \times \mathcal{K}^+ \times H \to \mathbb{R}, \quad L(\xi, \lambda; h) := \mathcal{G}(\xi, h) + (\lambda, \phi(\xi, h)), \tag{2.1}$$

where $\mathcal{K}^+ := \{\lambda \in Y^* \mid (\lambda, \xi) \le 0 \quad \text{for all } \xi \in \mathcal{K}\}$ is the cone polar to $\mathcal{K}$.

The first order optimality system for $(\mathrm{P})_h$ can be written in the form

$$\begin{aligned} &D_\xi L(\xi, \lambda; h) := D_\xi \mathcal{G}(\xi, h) + D_\xi \phi^*(\xi, h)\lambda = 0, \\ &\phi(\xi, h) \in N_{\mathcal{K}^+}(\lambda), \end{aligned} \tag{2.2}$$

where

$$N_{\mathcal{K}^+}(\lambda) = \left\{ y \in Y \;\middle|\; \begin{cases} (\mu - \lambda, y) \le 0 \quad \forall \mu \in \mathcal{K}^+ & \text{if } \lambda \in \mathcal{K}^+ \\ \emptyset & \text{if } \lambda \notin \mathcal{K}^+ \end{cases} \right\} \tag{2.3}$$

is the normal cone to $\mathcal{K}^+$. For the sake of simplicity, we denote

$$\begin{cases} \mathcal{F}: X \times Y^* \times H \to X^* \times Y, & \mathcal{T}: Y^* \to 2^{X^* \times Y}, \\ \mathcal{F}(\xi, \lambda, h) = \begin{bmatrix} D_\xi L(\xi, \lambda; h) \\ \phi(\xi, h) \end{bmatrix}, & \mathcal{T}(\lambda) = \begin{bmatrix} 0 \\ N_{\mathcal{K}^+} \end{bmatrix}. \end{cases} \tag{2.4}$$

Then (2.2) can be written in the form of the following inclusion (generalized equation):

$$\mathcal{F}(\xi, \lambda, h) \in \mathcal{T}(\lambda). \tag{2.5}$$

Let $\Xi \subset X$ and $\Lambda \subset Y^*$ be some closed and convex subsets. These subsets, endowed with the metric of X and Y^*, respectively, can be treated as complete nonlinear metric spaces (see Lemma 2.1 in [3]). In applications, Ξ and Λ are chosen as some sets of sufficiently regular functions. We will denote $\mathcal{B}_\tau^{\Xi \times \Lambda}(\cdot, \cdot) := (\Xi \times \Lambda) \cap \mathcal{B}_\tau^{X \times Y^*}(\cdot, \cdot)$.

In addition to (B1) we assume the following:

(B2) There exist closed convex sets $\Xi \subset X$ and $\Lambda \subset Y^*$ such that, for a fixed reference value $\widehat{h}$ of the parameter, there exists a pair $(\widehat{\xi}, \widehat{\lambda}) \in \Xi \times \Lambda$, which satisfies (2.2).

(B3) For any $h \in H$, $\mathcal{G}(\cdot, h)$ and $\phi(\cdot, h)$ are twice Fréchet differentiable on Ξ.

(B4) For any $\tau > 0$, there exists $k > 0$ such that

$$\begin{aligned} &\|\mathcal{F}(\xi, \lambda, h') - \mathcal{F}(\xi, \lambda, h'')\|_{X^* \times Y} \le k \|h' - h''\|_Z \\ &\qquad \forall\, (\xi, \lambda) \in \mathcal{B}_\tau^{\Xi \times \Lambda}(\widehat{\xi}, \widehat{\lambda}) \text{ and all } h', h'' \in H. \end{aligned} \tag{2.6}$$

Note that in (B4) both spaces Z and H are involved. We require that the estimate (2.6) hold for $h', h'' \in H$, whereas, on the right-hand side of this estimate, we have the norm of the space Z, which may be weaker than that of H.

Let us introduce the following *perturbed linearization* of (2.5):

$$\mathcal{F}(\widehat{\xi},\widehat{\lambda},\widehat{h})+D_\xi\mathcal{F}(\widehat{\xi},\widehat{\lambda},\widehat{h})(\eta-\widehat{\xi})+D_\lambda\mathcal{F}(\widehat{\xi},\widehat{\lambda},\widehat{h})(\kappa-\widehat{\lambda})-\delta\in\mathcal{T}(\kappa), \tag{2.7}$$

where $\delta:=(\delta_1,\delta_2)\in X^*\times Y$ is a perturbation. Note that, for $\delta=0$, $(\eta_0,\kappa_0)=(\widehat{\xi},\widehat{\lambda})$ is a solution of (2.7).

In view of (2.4), inclusion (2.7) can be interpreted as the optimality system for the following *accessory optimization problem*, depending on the perturbation δ:

$$(\mathrm{AP})_\delta \qquad \begin{array}{l}\min_{\eta\in X}\left\{\dfrac{1}{2}(\eta,D^2_{\xi\xi}L(\widehat{\xi},\widehat{\lambda};\widehat{h})\eta)-(\widehat{\delta}_1+\delta_1,\eta)\right\}\\ \text{subject to } D_\xi\phi(\widehat{\xi},\widehat{h})\eta-(\widehat{\delta}_2+\delta_2)\in\mathcal{K},\end{array}$$

where

$$\widehat{\delta}_1=D^2_{\xi\xi}L(\widehat{\xi},\widehat{\lambda};\widehat{h})\widehat{\xi}-D_\xi F(\widehat{\xi};\widehat{h}),\quad \widehat{\delta}_2=D_\xi\phi(\widehat{\xi},\widehat{h})\widehat{\xi}-\phi(\widehat{\xi};\widehat{h}). \tag{2.8}$$

Let us define the following function:

$$m(\xi,\lambda,h)=\mathcal{F}(\widehat{\xi},\widehat{\lambda},\widehat{h})+D_\xi\mathcal{F}(\widehat{\xi},\widehat{\lambda},\widehat{h})(\xi-\widehat{\xi})+D_\lambda\mathcal{F}(\widehat{\xi},\widehat{\lambda},\widehat{h})(\lambda-\widehat{\lambda})-\mathcal{F}(\xi,\lambda,h). \tag{2.9}$$

Our last assumptions are as follows:

(B5) There exist constants $\theta>0,\tau'>0,\pi'>0$, and $l>0$, as well as a subset $\Delta\subset X^*\times Y$ such that for each $\delta\in\mathcal{B}^\Delta_\theta(0):=\Delta\cap\mathcal{B}^{X^*\times Y}_\theta(0)$ there exists a unique stationary point $(\eta_\delta,\kappa_\delta)\in\mathcal{B}^{\Xi\times\Lambda}_{\tau'}(\widehat{\xi},\widehat{\lambda})$ of $(\mathrm{AP})_\delta$ and

$$\|\eta_{\delta'}-\eta_{\delta''}\|_X,\ \|\kappa_{\delta'}-\kappa_{\delta''}\|_{Y^*}\le l\|\delta'-\delta''\|_{X^*\times Y}\quad\forall\delta',\delta''\in\mathcal{B}^\Delta_\theta(0).$$

(B6) The inclusion $m(\xi,\lambda,h)\in\Delta$ is satisfied for all $(\xi,\lambda)\in\mathcal{B}^{\Xi\times\Lambda}_{\tau'}(\widehat{\xi},\widehat{\lambda})$ and all $h\in\mathcal{B}^H_{\pi'}(\widehat{h})$.

The following theorem will be the main tool in the stability analysis for nonlinear optimal control problems.

THEOREM 2.1. *Assume that* (B1)–(B6) *are satisfied. Then, there exist constants* $\pi>0$, $\tau>0$, *and* $\ell>0$ *such that for each* $h\in\mathcal{B}^H_\pi(\widehat{h})$ *there is a stationary point* (ξ_h,λ_h) *of* $(\mathrm{P})_h$, *unique in* $\mathcal{B}^{\Xi\times\Lambda}_\tau(\widehat{\xi},\widehat{\lambda})$, *and*

$$\|\xi_{h'}-\xi_{h''}\|_X,\ \|\lambda_{h'}-\lambda_{h''}\|_{Y^*}\le\ell\|h'-h''\|_H\quad\forall h',h''\in\mathcal{B}^H_\pi(\widehat{h}). \tag{2.10}$$

Remark 2.2. Theorem 2.1 is a slight modification of Robinson's implicit function theorem for strongly regular generalized equations (see Theorem 2.1 in [16]). The difference is that, in Theorem 2.1, there are considered subsets $\Delta\subset X^*\times Y$ and $\Xi\times\Lambda\subset X\times Y^*$ rather than the whole spaces. In applications to optimal control problems, this modification allows us to overcome the difficulty connected with the so-called *two-norm discrepancy* [13] by exploiting the regularity of the stationary points. The proofs of Theorem 2.1 are based on some modifications of the original Robinson's proof and they can be found in [7] (Theorem 2.2) and in [3] (Lemma 2.1).

3. Optimal control problem. In this section our model optimal control problem is formulated and basic assumptions are introduced. Let

$$\left\{\begin{array}{ll}H & =W^{2,\infty}(0,1;\mathbb{R}):=\{h\in L^\infty(0,1;\mathbb{R})\mid\ddot{h}\in L^\infty(0,1;\mathbb{R})\}\quad\text{and}\\ X^p & =W^{1,p}_0(0,1;\mathbb{R}^n)\times L^p(0,1;\mathbb{R}^m),\quad p\in[1,\infty],\end{array}\right. \tag{3.1}$$

be the spaces of parameters and arguments, respectively.

Consider the family of the following optimal control problems depending on $h \in H$:

$(\mathrm{O})_h$ Find $(x_h, u_h) \in X^2$ such that

$$F(x_h, u_h, h) = \min \left\{ F(x,u,h) := \int_0^1 \varphi(x(t), u(t), h(t))dt \right\} \tag{3.2}$$

subject to

$$\dot{x}(t) - f(x(t), u(t), h(t)) = 0 \quad \text{for almost all } t \in [0,1], \tag{3.3}$$

$$x(0) = 0, \tag{3.4}$$

$$\vartheta(x(t), h(t)) \le 0 \qquad \forall\, t \in [0,1], \tag{3.5}$$

where $\varphi:\ \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R} \to \mathbb{R},\ f:\ \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}^n,\ \vartheta:\ \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$.

Remark 3.1. To minimize technicalities, we consider the simple homogeneous initial condition. However, the same approach can be applied to general two-point boundary value problems. Also vector-valued state constraints can be considered and additional control constraints can be included.

The following standing assumptions are assumed to be satisfied throughout the paper:

(I) There exist open sets $\mathcal{R}^n \subset \mathbb{R}^n$ and $\mathcal{R}^m \subset \mathbb{R}^m$ such that the functions $\varphi(\cdot,\cdot,\cdot)$, $D_x\varphi(\cdot,\cdot,\cdot)$, and $D_u\varphi(\cdot,\cdot,\cdot)$ as well as $f(\cdot,\cdot,\cdot)$, $D_x f(\cdot,\cdot,\cdot)$, and $D_u f(\cdot,\cdot,\cdot)$ are Fréchet differentiable in (x,u,h) on $\mathcal{R}^n \times \mathcal{R}^m \times \mathbb{R}$. The functions $\vartheta(\cdot,\cdot)$ and $D_x\vartheta(\cdot,\cdot)$ are twice Fréchet differentiable in (x,h) on $\mathcal{R}^n \times \mathbb{R}$.

(II) For a given reference value $\widehat{h} \in H$ of the parameter there exists a reference solution $(\widehat{x}, \widehat{u})$ of $(\mathrm{O})_{\widehat{h}}$, where $\widehat{u} \in C(0,1;\mathbb{R}^m)$ and $(\widehat{x}(t), \widehat{u}(t)) \in \mathcal{R}^n \times \mathcal{R}^m$ for all $t \in [0,1]$.

To simplify notation, the functions evaluated at the reference solution will be denoted by a hat, e.g., $\widehat{\varphi} := \varphi(\widehat{x}, \widehat{u}, \widehat{h})$, $\widehat{\vartheta} := \vartheta(\widehat{x}, \widehat{h})$. Let us define the spaces of multipliers

$$Y^p = L^p(0,1;\mathbb{R}^n) \times W^{1,p}(0,1;\mathbb{R}), \qquad p \in [1,\infty], \tag{3.6}$$

and introduce the Lagrangian and Hamiltonian for $(\mathrm{O})_h$:

$$\mathcal{L} : X^2 \times Y^2 \times H \to \mathbb{R}, \quad \mathcal{H} : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R},$$

$$\begin{cases} \mathcal{L}(x,u,p,\mu;h) = F(x,u,h) - (p, \dot{x} - f(x,u,h)) \\ \qquad +\mu(0)\vartheta(x(0),h(0)) + (\dot{\mu}, D_x\vartheta(x,h)f(x,u,h) + D_h\vartheta(x,h)\dot{h}), \end{cases} \tag{3.7}$$

$$\begin{cases} \mathcal{H}(x(t),u(t),p(t),\dot{\mu}(t);h(t),\dot{h}(t)) = \varphi(x(t),u(t)) + \langle p(t), f(x(t),u(t),h(t))\rangle \\ \qquad +\dot{\mu}(t)(D_x\vartheta(x(t),h(t))f(x(t),u(t),h(t)) + D_h\vartheta(x(t),h(t))\dot{h}(t)). \end{cases} \tag{3.8}$$

Remark 3.2. Lagrangian (3.7) is *normal*; i.e., the Lagrange multiplier corresponding to the functional $F(x,u,h)$ is different from zero. The Lagrangian is in the so-called *indirect* or Pontryagin form, with the absolute continuous adjoint variable (see section 5 in [6], as well as [5] and [14]). The state constraints are considered

in the space $W^{1,2}(0,1;\mathbb{R})$, where the general form of a linear functional is given by $\mu(0)y(0)+(\dot\mu,\dot y)$, with $\mu\in W^{1,2}(0,1;\mathbb{R})$. Hence, using the state equation, we get

$$\begin{aligned}&\mu(0)\vartheta(x(0),h(0))+\left(\dot\mu,\frac{d}{dt}\vartheta(x,h)\right)\\&\quad=\mu(0)\vartheta(x(0),h(0))+(\dot\mu,D_x\vartheta(x,h)f(x,u,h)+D_h\vartheta(x,h)\dot h).\end{aligned}$$

It turns out that, in stability analysis, the Lagrangian in indirect form is more convenient than that in direct form, thanks to the regularity of the Lagrange multiplier μ.

Denote by $K=\{d\in W^{1,2}(0,1;\mathbb{R})\mid d(t)\le 0\}$ the cone of nonpositive functions in $W^{1,2}(0,1;\mathbb{R})$. The cone polar to K is given (see, e.g., [15]) by

$$K^+=\left\{\mu\in W^{1,2}(0,1;\mathbb{R})\;\middle|\;\left\{\begin{array}{l}\mu(0)-\dot\mu(0+)\ge 0,\ \dot\mu(t)\ge 0\\ \text{and }\dot\mu(\cdot)\text{ is nonincreasing}\end{array}\right.\right\}. \tag{3.9}$$

Clearly, if $\mu\in W^{2,2}(0,1;\mathbb{R})$, the last condition in (3.9) reduces to $\ddot\mu^i(t)\le 0$ for almost all $t\in[0,1]$. By

$$\mathcal{N}_{K^+}(\mu):=\left\{\begin{array}{ll}\{y\in W^{1,2}(0,1;\mathbb{R})\mid (y,\nu-\mu)_{1.2}\le 0\ \forall\nu\in K^+\}&\text{if }\mu\in K^+,\\ \emptyset&\text{if }\mu\notin K^+,\end{array}\right. \tag{3.10}$$

we denote the normal cone to K^+ at μ.

The stationary conditions of Lagrangian (3.7) can be expressed by the following system:

$$\left\{\begin{array}{l}\dot p+D_x\mathcal{H}(x,u,p,\dot\mu;h,\dot h)\\ \quad=\dot p+D_xf^*(x,u,h)p+D_x\varphi(x,u,h)+(D_xf^*(x,u,h)D_x\vartheta^*(x,h)\\ \qquad+D^2_{xx}\vartheta(x,h)f(x,u,h)+D^2_{hx}\vartheta(x,h)\dot h)\dot\mu=0,\\ p(1)=0,\end{array}\right. \tag{3.11}$$

$$\left\{\begin{array}{l}D_u\mathcal{H}(x,u,p,\dot\mu;h,\dot h)\\ \quad=D_u\varphi(x,u,h)+D_uf^*(x,u,h)p+D_uf^*(x,u,h)D_x\vartheta^*(x,h)\dot\mu=0,\end{array}\right. \tag{3.12}$$

$$\vartheta(x,h)\in\mathcal{N}_{K^+}(\mu). \tag{3.13}$$

Note that condition (3.13) is equivalent to the standard Karush–Kuhn–Tucker sign and complementarity conditions for the inequality constraints. For the sake of simplicity, we will denote $\xi=(x,u)\in X^2$, $\lambda=(p,\mu)\in Y^2$.

The purpose of this paper is to study the local properties of the map $H\ni h\mapsto(\zeta_h,\lambda_h)\in X^2\times Y^2$. More precisely, we are looking for conditions under which there exist a constant $\pi>0$ and a subset $\mathcal{Z}\subset X^2\times Y^2$, containing the reference point $(\widehat\zeta,\widehat\lambda)$ such that for each $h\in\mathcal{B}^H_\pi(\widehat h)$ there exists a unique stationary point $(\zeta_h,\lambda_h)\in\mathcal{Z}$, which is a Lipschitz continuous function of h.

To cope with this problem, we will need several assumptions to be satisfied at the reference point. These assumptions consist of constraint qualifications and coercivity conditions. To formulate constraint qualifications, for a fixed $\alpha\ge 0$ we introduce the sets of α-active constraints:

$$M_\alpha=\{t\in[0,1]\mid\vartheta(\widehat x(t),\widehat h(t))\ge-\alpha\}. \tag{3.14}$$

Assume the following:

(H1) There exist $\alpha>0$ such that $0\notin M_\alpha$.

(H2) (*linear independence*). There exist $\alpha > 0$ and $\chi > 0$ such that

$$|D_u \widehat{f}^*(t) D_x \widehat{\vartheta}^*(t)| \geq \chi \quad \forall\, t \in M_\alpha.$$

Note that by (H2) the analysis is restricted to the so-called *first order state constraints* [6].

It will be more convenient to modify assumption (H2). To this end, for $\alpha \geq 0$ denote

$$T_\alpha(t) = \min\{\widehat{\vartheta}(t) + \alpha, 0\}. \tag{3.15}$$

It can be easily shown (see Lemma 2.1 in [9]) that (H2) is equivalent to the following condition:

(H2′) There exists $\alpha' > 0$ and $\chi' > 0$ such that

$$\left| \begin{array}{c} D_u \widehat{f}^*(t) D_x \widehat{\vartheta}^*(t) \\ T_{\alpha'}(t) \end{array} \right| \geq \chi' \quad \forall\, t \in [0,1].$$

Let us introduce the map

$$\mathcal{C}_\alpha :\ X^2 \times W^{1,2}(0,1;\mathbb{R}) \to L^2(0,1;\mathbb{R}^n) \times W^{1,2}(0,1;\mathbb{R}),$$

$$\mathcal{C}_\alpha \begin{bmatrix} y \\ v \\ \kappa \end{bmatrix} = \begin{bmatrix} \dot{y} - D_x \widehat{f} y - D_u \widehat{f} v \\ D_x \widehat{\vartheta} y + T_\alpha \kappa \end{bmatrix}. \tag{3.16}$$

By Lemma 4.1 and Theorem 4.3 in [10], we obtain the following.

LEMMA 3.3. *If assumptions* (H1) *and* (H2′) *are satisfied, then the map* $\mathcal{C}_{\alpha'}$ *is surjective.*

LEMMA 3.4. *If assumptions* (H1) *and* (H2′) *are satisfied, then there exists a unique Lagrange multiplier* $\widehat{\lambda} = (\widehat{p}, \widehat{\mu}) \in Y^2$ *such that the first order optimality conditions* (3.11)–(3.13) *hold at* $(\widehat{x}, \widehat{u}, \widehat{p}, \widehat{\mu})$.

In addition to the constraint qualifications, we will need some coercivity conditions. Assume the following:

(H3) (*Legendre–Clebsch condition*). There exists $\bar{\gamma} > 0$ such that

$$\langle v, D^2_{uu} \widehat{\mathcal{L}}\, v \rangle \geq \bar{\gamma} |v|^2 \qquad \forall\, v \in \mathbb{R}^m \text{ and all } t \in [0,1].$$

The following regularity result follows from Theorem 2.1 in [5] (see Proposition 6.6 in [8]).

LEMMA 3.5. *If assumptions* (H1)–(H3) *are satisfied, then* $\dot{\widehat{x}}, \widehat{u}, \dot{\widehat{p}}, \dot{\widehat{\mu}}$ *are Lipschitz continuous on* $[0,1]$*, with the Lipschitz modulus denoted by* $\widehat{\varsigma} > 0$.

In view of the uniqueness and regularity of $\widehat{\mu}$, we can introduce the following sets, depending on the parameter $\alpha > 0$:

$$N_\alpha = [0,1] \setminus \overline{\{t \in [0,1] \mid -\ddot{\widehat{\mu}}(t) \leq \alpha\}}, \quad \text{as well as } N_0 = \bigcup_{\alpha > 0} N_\alpha. \tag{3.17}$$

The sets N_α are open in $[0,1]$. Define the following subspace of X^2:

$$\mathcal{E}_\alpha = \left\{ (y,v) \in X^2 \;\middle|\; \left\{ \begin{array}{l} \dot{y}(t) - D_x \widehat{f}(t) y(t) - D_u \widehat{f}(t) v(t) = 0, \\ \langle D_x \widehat{\vartheta}(t), y(t) \rangle = 0 \ \forall\, t \in N_\alpha, \\ \langle D_x \widehat{\vartheta}(1), y(1) \rangle = 0 \ \text{ if } \dot{\widehat{\mu}}(1) > 0. \end{array} \right. \right\} \tag{3.18}$$

For the sake of simplicity we will denote $D^2\widehat{\mathcal{L}} := D^2_{(x,u)(x,u)}\mathcal{L}\,(\widehat{x},\widehat{u},\widehat{p},\widehat{\mu};\widehat{h})$.

Assume the following:

(H4) (*coercivity*). There exist constants $\alpha > 0$ and $\gamma > 0$ such that

$$\Big((y,v), D^2\widehat{\mathcal{L}}\,(y,v)\Big) \geq \gamma(\|y\|^2_{1,2} + \|v\|^2_2) \quad \forall\ (y,v) \in \mathcal{E}_\alpha. \tag{3.19}$$

Remark 3.6. The coercivity condition (H4) takes into account strongly active state constraints. It is weaker than the strong coercivity condition, where the active inequality constraints are ignored. The latter condition was used in stability analysis in [8] and in [3]. The application of the weaker condition (H4) is the main new contribution of this paper. It turns out that the weakening of the coercivity condition is crucial. Namely, as it will be shown in a forthcoming paper of the author, conditions of the form (H1)–(H4) are not only sufficient, but also necessary, for Lipschitz stability and directional differentiability of the solutions and Lagrange multipliers for a class of state-constrained optimal control problems. Thus, they constitute a characterization of these properties.

The following result is a slight modification of Theorem 4.1 in [1].

LEMMA 3.7. *Suppose that* (H1)–(H3), *as well as* (H4) *with* $\alpha = 0$, *hold. Then there exist* $\rho > 0$ *and* $c > 0$ *such that*

$$\begin{aligned} F(x,u,\widehat{h}) - F(\widehat{x},\widehat{u},\widehat{h}) \geq c(\|x-\widehat{x}\|^2_{1,2} + \|u-\widehat{u}\|^2_2), \\ \forall\ (x,u) \in \mathcal{B}^{X^\infty}_\rho(\widehat{x},\widehat{u}) \text{ feasible for } (\mathrm{O})_{\widehat{h}}. \end{aligned} \tag{3.20}$$

Thus $(\widehat{x},\widehat{u})$ *is a second order local minimizer of* $(\mathrm{O})_{\widehat{h}}$.

For the sake of completeness, the proof of Lemma 3.7 is given in Appendix A.

4. Stability analysis. In this section the abstract Theorem 2.1 will be applied to the optimal control problems $(\mathrm{O})_h$. To this end, we have to express $(\mathrm{O})_h$ in terms of the cone-constrained problem $(\mathrm{P})_h$. We set

$$\left\{\begin{array}{l} Z = W^{1,2}(0,1;\mathbb{R}), \quad H = W^{2,\infty}(0,1;\mathbb{R}), \quad X = X^2, \quad Y = Y^2, \\ (X^2)^* = L^2(0,1;\mathbb{R}^n) \times L^2(0,1;\mathbb{R}^m), \quad (Y^2)^* = Y^2, \\ \xi = (x,u), \quad \lambda = (p,\mu), \qquad \mathcal{K} = \{0\} \times K, \\ \\ \mathcal{G}(\xi,h) = F(x,u,h), \quad \phi(\xi,h) = \left[\begin{array}{c} \dot{x} - f(x,u,h) \\ \vartheta(x,h) \end{array}\right]. \end{array}\right. \tag{4.1}$$

It can be easily checked that by (I), condition (B1) is satisfied. In order to verify the remaining assumptions of Theorem 2.1, we have to introduce the sets Ξ and Λ needed in (B2)–(B6). We define

$$\Xi = \{(x,u) \in X \mid \|\ddot{x}\|_\infty,\ \|\dot{u}\|_\infty \leq \varsigma\}, \quad \Lambda = \{(p,\mu) \in Y^* \mid \|\ddot{p}\|_\infty,\ \|\ddot{\mu}\|_\infty \leq \varsigma\}, \tag{4.2}$$

where the constant $\varsigma > \widehat{\varsigma}$ will be given later.

It can be easily seen that the sets Ξ and Λ are convex and closed, and they can be treated as complete nonlinear metric spaces (see [3]). Thus in view of (II) and Lemma 3.5, condition (B2) is satisfied. Moreover, by (I), $\mathcal{G}(\cdot,h)$ and $\phi(\cdot,h)$ are twice Fréchet differentiable on Ξ, for any $h \in H$. So condition (B3) holds. Similarly, (B4) is satisfied.

To verify (B5), we have to construct the accessory problem $(\mathrm{AO})_\delta$ for $(\mathrm{O})_h$. To this end, we introduce perturbations

$$\left\{\begin{array}{l} \delta = (\delta_1,\delta_2,\delta_3,\delta_4) \in (X^2)^* \times Y^2, \qquad \text{where} \\ (\delta_1,\delta_2,\delta_3,\delta_4) \in L^2(0,1;\mathbb{R}^n) \times L^2(0,1;\mathbb{R}^m), \times L^2(0,1;\mathbb{R}^n) \times W^{1,2}(0,1;\mathbb{R}). \end{array}\right. \tag{4.3}$$

The accessory problem takes the form of the following linear-quadratic optimal control:

$$
\begin{aligned}
(\mathrm{AO})_\delta \qquad & \text{Find } \eta_\delta := (y_\delta, v_\delta) \in X^2 \text{ such that} \\
& J(y_\delta, v_\delta; \delta) = \min J(y, v; \delta) \quad \text{subject to} \\
& \dot{y}(t) - D_x \widehat{f}(t) y(t) - D_u \widehat{f}(t) v(t) - (\widehat{\delta}_3 + \delta_3(t)) = 0, \\
& D_x \widehat{\vartheta}(t) y(t) - (\widehat{\delta}_4(t) + \delta_4(t)) \le 0 \quad \forall\, t \in [0,1],
\end{aligned}
$$

where

$$
\begin{aligned}
& J(y, v; \delta) = \frac{1}{2}((y,v), D^2 \widehat{\mathcal{L}}\,(y,v)) - (\widehat{\delta}_1 + \delta_1, y) - (\widehat{\delta}_2 + \delta_2, v) \\
& \widehat{\delta}_1(t) = D^2_{xx} \widehat{\mathcal{H}}(t) \widehat{x}(t) + D^2_{xu} \widehat{\mathcal{H}}(t) \widehat{u}(t) - D_x \widehat{\varphi}(t), \\
& \widehat{\delta}_2(t) = D^2_{ux} \widehat{\mathcal{H}}(t) \widehat{x}(t) + D^2_{uu} \widehat{\mathcal{H}}(t) \widehat{u}(t) - D_u \widehat{\varphi}(t), \\
& \widehat{\delta}_3(t) = \widehat{f}(t) - D_x \widehat{f}(t) \widehat{x}(t) - D_u \widehat{f}(t) \widehat{u}(t), \\
& \widehat{\delta}_4(t) = D_x \widehat{\vartheta}(t) \widehat{x}(t) - \widehat{\vartheta}(t).
\end{aligned} \tag{4.4}
$$

Note that, in view of (I) and Lemma 3.5, there exists a constant $s > 0$ such that

$$
\|\dot{\widehat{\delta}}_1\|_\infty,\ \|\dot{\widehat{\delta}}_2\|_\infty,\ \|\dot{\widehat{\delta}}_3\|_\infty,\ \|\ddot{\widehat{\delta}}_4\|_\infty \le s. \tag{4.5}
$$

The set Δ needed in (B5)–(B6) is defined as

$$
\Delta = \{(\delta_1, \delta_2, \delta_3, \delta_4) \in X^* \times Y \mid \|\dot{\delta}_1\|_\infty, \|\dot{\delta}_2\|_\infty, \|\dot{\delta}_3\|_\infty, \|\ddot{\delta}_4\|_\infty \le s\}, \tag{4.6}
$$

where $s > 0$ is given in (4.5). To verify (B5), we have to show the existence, stability, and regularity of the stationary points of $(\mathrm{AO})_\delta$, for $\delta \in \Delta$.

To get the needed existence and stability results for $(\mathrm{AO})_\delta$, we will use two important lemmas. To formulate them, for any open set $D \subset [0,1]$, introduce the following superspace of the subspace $\mathcal{E}_\alpha$ defined in (3.18):

$$
\mathcal{E}_{\alpha,D} = \left\{ (y,v) \in X^2 \;\middle|\; \left\{ \begin{array}{l} \dot{y}(t) - D_x \widehat{f}(t) y(t) - D_u \widehat{f}(t) v(t) = 0, \\ \langle D_x \widehat{\vartheta}(t), y(t) \rangle = 0 \ \forall\, t \in N_\alpha \setminus D, \\ \langle D_x \widehat{\vartheta}(1), y(1) \rangle = 0 \ \text{ if } \dot{\widehat{\mu}}(1) > 0. \end{array} \right. \right\} \tag{4.7}
$$

Lemma 4.1. *Suppose that* (H1)–(H4) *hold. There exist constant* $\beta > 0$ *and* $\widetilde{\gamma} > 0$ *such that, if* $N_\alpha \cap D$ *does not contain any subinterval* (t', t'') *of the length larger than* β*, that is, if*

$$
|t'' - t'| \le \beta \quad \forall\, (t', t'') \in N_\alpha \cap D, \tag{4.8}
$$

then

$$
\Big((y,v), D^2 \widehat{\mathcal{L}}\,(y,v)\Big) \ge \widetilde{\gamma}(\|y\|^2_{1,2} + \|v\|^2_2) \quad \forall\, (y,v) \in \mathcal{E}_{\alpha,D}. \tag{4.9}
$$

The second lemma concerns a property of elements of $W^{1,2}_0(0,1;\mathbb{R})$, in a neighborhood of $\widehat{\mu}$. For an arbitrary $\nu \in W^{1,2}_0(0,1;\mathbb{R})$ denote

$$
D^\nu = \{t \in [0,1] |\ \dot{\nu}(\cdot) = \text{const a.e. in a neighborhood of } t\}. \tag{4.10}
$$

LEMMA 4.2. *Choose any $\eta > 0$ and let $\nu \in W_0^{1,2}(0,1;\mathbb{R})$ be such that*

$$\|\nu - \widehat{\mu}\|_{1,2} \leq \eta.$$

Let $(t', t'') \subset N_\alpha \cap D^\nu$ be any subinterval belonging to $N_\alpha \cap D^\nu$. Then

$$\text{meas}\,(t', t'') \leq \left(12\left(\frac{\eta}{\alpha}\right)^2\right)^{1/3}. \tag{4.11}$$

Using Lemmas 4.1 and 4.2, we get the following.

LEMMA 4.3. *If assumptions* (H1)–(H4) *are satisfied, then there exist constants $\theta > 0$, $\tau > 0$, and $r > 0$ such that, for each $\delta \in \mathcal{B}_\theta^{X^* \times Y}(0)$, there is a unique stationary point $(y_\delta, v_\delta, q_\delta, \nu_\delta) \in \mathcal{B}_\tau^{X^2 \times Y^2}(\widehat{\xi}, \widehat{\lambda})$ of* $(\text{AO})_\delta$, *and*

$$\begin{aligned} &\|y_\delta - \widehat{x}\|_{1,2}, \|v_\delta - \widehat{u}\|_2, \|q_\delta - \widehat{p}\|_{1,2}, \|\nu_\delta - \widehat{\mu}\|_{1,2} \leq r\|\delta\|_{X^* \times Y} \\ &\qquad \forall\ \delta \in \mathcal{B}_\theta^{X^* \times Y}(0). \end{aligned} \tag{4.12}$$

Moreover, there exists $\varsigma > \widehat{\varsigma}$ such that

$$\|\ddot{y}_\delta\|_\infty, \|\dot{v}_\delta\|_\infty, \|\ddot{q}_\delta\|_\infty, \|\ddot{\nu}_\delta\|_\infty \leq \varsigma \quad \forall\ \delta \in \mathcal{B}_\theta^\Delta(0). \tag{4.13}$$

In definition (4.2) of Ξ and Λ, we choose ς given in Lemma 4.3. Thus we get

$$(y_\delta, v_\delta) \in \Xi, \quad (q_\delta, v_\delta) \in \Lambda \quad \forall\ \delta \in \mathcal{B}_\theta^\Delta(0).$$

Using Lemma 4.3 together with Lemmas 4.1 and 4.2, we arrive at the following stability result for $(\text{AO})_\delta$.

PROPOSITION 4.4. *If assumptions* (H1)–(H4) *are satisfied, then there exist constants $\theta > 0$, $\tau > 0$, and $l > 0$ such that, for each $\delta \in \mathcal{B}_\theta^{X^* \times Y}(0)$, there is a unique stationary point $(y_\delta, v_\delta, q_\delta, \nu_\delta) \in \mathcal{B}_\tau^{\Xi \times \Lambda}(\widehat{\xi}, \widehat{\lambda})$ of* $(\text{AO})_\delta$, *and*

$$\begin{aligned} &\|y_{\delta'} - y_{\delta''}\|_{1,2}, \|v_{\delta'} - v_{\delta''}\|_2, \|q_{\delta'} - q_{\delta''}\|_{1,2}, \|\nu_{\delta'} - \nu_{\delta''}\|_{1,2} \\ &\qquad \leq l\|\delta' - \delta''\|_{X^* \times Y} \quad \forall\ \delta', \delta'' \in \mathcal{B}_\theta^\Delta(0). \end{aligned} \tag{4.14}$$

By Proposition 4.4, condition (B5) holds. To apply Theorem 2.1, we still have to verify (B6). The function m, given in (2.9), can be rewritten as follows:

$$\begin{aligned} m(\xi, \lambda, h) \ &= \int_0^1 \left[D_\xi \mathcal{F}(\widehat{\xi}, \widehat{\lambda}, \widehat{h}) - D_\xi \mathcal{F}(\xi_\beta, \lambda_\beta, \widehat{h})\right] d\beta\ (\xi - \widehat{\xi}) \\ &+ \int_0^1 \left[D_\lambda \mathcal{F}(\widehat{\xi}, \widehat{\lambda}, \widehat{h}) - D_\lambda \mathcal{F}(\xi_\beta, \lambda_\beta, \widehat{h})\right] d\beta\ (\lambda - \widehat{\lambda}) \\ &- \int_0^1 D_h \mathcal{F}(\xi, \lambda, h_\beta) d\beta\ (h - \widehat{h}), \end{aligned} \tag{4.15}$$

where $\mathcal{F}$ is as defined in (2.4), with $\mathcal{G}$ and ϕ as given in (4.1), whereas $\xi_\beta = (1-\beta)\widehat{\xi} + \beta\xi$, $\lambda_\beta = (1-\beta)\widehat{\lambda} + \beta\lambda$, $h_\beta = (1-\beta)\widehat{h} + \beta h$. By straightforward calculations, we get the following result, which shows that assumption (B6) holds.

LEMMA 4.5. *If assumptions* (H1)–(H4) *are satisfied, then there exist $\pi' > 0$ and $\tau > 0$ such that $m(\xi, \lambda, h)$ belongs to Δ, for any $(\xi, \lambda) \in \mathcal{B}_\tau^{\Xi \times \Lambda}(\widehat{\xi}, \widehat{\lambda})$ and any $h \in \mathcal{B}_{\pi'}^H(\widehat{h})$.*

Thus, all assumptions of Theorem 2.1 hold and, by that theorem, we obtain the following principal stability result of this paper.

THEOREM 4.6. *If assumptions* (H1)–(H4) *are satisfied, then there exist constants* $\pi > 0$, $\tau > 0$, *and* $\ell > 0$ *such that, for each* $h \in \mathcal{B}_\pi^H(\widehat{h})$, *there is a unique stationary point* $(\xi_h, \lambda_h) := (x_h, u_h, p_h, \mu_h) \in \mathcal{B}_\tau^{\Xi\times\Lambda}(\widehat{\xi}, \widehat{\lambda})$ *of* $(\mathrm{O})_h$ *and*

$$\begin{aligned} &\|x_{h'} - x_{h''}\|_{1,2}, \|u_{h'} - u_{h''}\|_2, \|p_{h'} - p_{h''}\|_{1,2}, \|\mu_{h'} - \mu_{h''}\|_{1,2} \le \ell\|h' - h''\|_Z \\ &\qquad \forall\ h', h'' \in \mathcal{B}_\pi^H(\widehat{h}). \end{aligned} \tag{4.16}$$

In view of the regularity (4.2) of (x_h, u_h, p_h, μ_h), the estimate (4.16) implies (see Lemma 3.1 in [3]) the following.

COROLLARY 4.7. *If the assumptions of Theorem* 4.6 *hold, then there exist* $\pi > 0$ *and* $\ell_\infty > 0$ *such that*

$$\begin{aligned} &\|x_{h'} - x_{h''}\|_{1,\infty}, \|u_{h'} - u_{h''}\|_\infty, \|p_{h'} - p_{h''}\|_{1,\infty}, \|\mu_{h'} - \mu_{h''}\|_{1,\infty} \le \ell_\infty\|h' - h''\|_Z^{2/3} \\ &\qquad \forall\ h', h'' \in \mathcal{B}_\pi^H(\widehat{h}). \end{aligned} \tag{4.17}$$

Using Theorem 4.6 and Lemma 3.7, we show that (x_h, u_h) in Theorem 4.6 is a solution of $(\mathrm{O})_h$. Thus we get the following.

COROLLARY 4.8. *If assumptions* (H1)–(H4) *are satisfied, then, for* $\pi > 0$ *sufficiently small, the stationary point* (ξ_h, λ_h) *in Theorem* 4.6 *corresponds to the solution and Lagrange multiplier of* $(\mathrm{O})_h$.

5. Proofs. In this section we present the proofs of the stability results formulated in section 4.

Proof of Lemma 4.1. Suppose that the assertion of the lemma is not true. Then, for each $\beta_i > 0$ and $\gamma_i > 0$ there exist an open set $D_i \subset [0, 1]$ and a pair $(y_i, v_i) \in \mathcal{E}_{\alpha,D_i}$, with $\|v_i\|_2 = 1$ such that

$$|t'' - t'| \le \beta_i \quad \forall\ (t', t'') \subset N_\alpha \cap D_i \tag{5.1}$$

and

$$\Big((y_i, v_i), D^2\widehat{\mathcal{L}}\,(y_i, v_i)\Big) < \gamma_i(\|y_i\|_{1,2}^2 + \|v_i\|_2^2). \tag{5.2}$$

Let us choose a sequence $\{(\beta_i, \gamma_i)\} \to (0, 0)$ and let $\{(y_i, v_i)\}$ be the corresponding sequence of normalized pairs $(y_i, v_i) \in \mathcal{E}_{\alpha,D_i}$ satisfying (5.1) and (5.2). From $\{(y_i, v_i)\}$, we can extract a weakly convergent subsequence, still denoted by $\{(y_i, v_i)\}$. Thus, there exists $(\widetilde{y}, \widetilde{v}) \in X^2$ such that

$$\begin{cases} v_i \rightharpoonup \widetilde{v} & \text{weakly in } L^2(0, 1; \mathbb{R}^m), \\ y_i \rightharpoonup \widetilde{y} & \text{weakly in } W^{1,2}(0, 1; \mathbb{R}^n), \text{ i.e., strongly in } C(0, 1; \mathbb{R}^n). \end{cases} \tag{5.3}$$

Since

$$\Big((y, v), D^2\widehat{\mathcal{L}}\,(y, v)\Big) = (y, D_{xx}^2\widehat{\mathcal{L}}\,y) + 2(y, D_{xu}^2\widehat{\mathcal{L}}\,v) + (v, D_{uu}^2\widehat{\mathcal{L}}\,v), \tag{5.4}$$

by (5.3), the first two components on the right-hand side of (5.4) are continuous in the weak topology of X^2, whereas by assumption (H3), the third component is

weakly lower semicontinuous in $L^2(0,1;\mathbb{R}^m)$. Hence $((y,v), D^2\widehat{\mathcal{L}}\,(y,v))$ is weakly lower semicontinuous in X^2. Thus, from (5.2) we obtain

$$\begin{aligned}\Big((\widetilde{y},\widetilde{v}), D^2\widehat{\mathcal{L}}\,(\widetilde{y},\widetilde{v})\Big) &\le \liminf \Big((y_i,v_i), D^2\widehat{\mathcal{L}}\,(y_i,v_i)\Big)\\ &\le \limsup \Big((y_i,v_i), D^2\widehat{\mathcal{L}}\,(y_i,v_i)\Big) \le 0.\end{aligned} \tag{5.5}$$

On the other hand, from (5.1) and (5.3) we get $(\widetilde{y},\widetilde{v}) \in \mathcal{E}_\alpha$. Hence, in view of (H3), we obtain from (5.5) that $0 = ((\widetilde{y},\widetilde{v}), D^2\widehat{\mathcal{L}}\,(\widetilde{y},\widetilde{v}))$, which by (H4) implies $(\widetilde{y},\widetilde{v}) = (0,0)$. Thus, by (5.3) and (5.4), we get from (5.5) that $0 = (\widetilde{v}, D^2_{uu}\widehat{\mathcal{L}}\widetilde{v}) = \lim(v_i, D^2_{uu}\widehat{\mathcal{L}}v_i)$. By coercivity of $D^2_{uu}\widehat{\mathcal{L}}$, that is equivalent to $0 = \|\widetilde{v}\|_2^2 = \lim\|v_i\|_2^2$, which, together with (5.3) implies that $v_i \to \widetilde{v}$ strongly in $L^2(0,1;\mathbb{R}^m)$. Since $\|v_i\|_2 = 1$, the strong convergence contradicts the fact that $\widetilde{v} = 0$ and completes the proof. □

Proof of Lemma 4.2. Using definitions (3.17) and (4.10), we obtain

$$\begin{aligned}\eta^2 &\ge \|\nu - \widehat{\mu}\|_{1,2}^2 \ge \int_{t'}^{t''} (\dot{\nu}(t) - \dot{\widehat{\mu}}(t))^2 dt \ge \min_{c\in\mathbb{R}} \int_{t'}^{t''} [c - \alpha\, t]^2 dt\\ &\ge \frac{1}{12}\alpha^2\ (\text{meas}\ ([t',t'']))^3,\end{aligned}$$

which implies (4.11). □

Proof of Lemma 4.3. Denote by $\mathcal{D}_\beta$ the family of all open subsets D of $(0,1)$ such that (4.8) holds. For a fixed $D \in \mathcal{D}_\beta$, we introduce the following modification of problem $(\text{AO})_\delta$:

$$\begin{aligned}(\text{AO})^D_\delta \qquad &\text{Find } \eta^D_\delta := (x^D_\delta, u^D_\delta) \in X^2 \text{ such that}\\ &J(x^D_\delta, u^D_\delta; \delta) = \min J(y,v;\delta) \quad \text{subject to}\\ &\dot{y}(t) - D_x\widehat{f}(t)y(t) - D_u\widehat{f}(t)v(t) - (\widehat{\delta}_3(t) + \delta_3(t)) = 0,\\ &D_x\widehat{\vartheta}(t)y(t) - (\widehat{\delta}_4(t) + \delta_4(t)) \begin{cases} = 0 & \text{for } t \in N_\alpha \setminus D,\\ \le 0 & \text{for } t \in [0,1] \setminus (N_\alpha \setminus D).\end{cases}\end{aligned}$$

Note that $(\text{AO})^D_\delta$ differs from $(\text{AO})_\delta$ only in the form of the inequality constraints. In view of Lemma 4.1, the quadratic term in the cost functional $J(y,v,;\delta)$ of $(\text{AO})^D_\delta$ satisfies the coercivity condition (4.9) on the linear hull of the feasible set. Stability of coercive linear-quadratic problems was studied in, among others, [3]. By Lemmas 3.8 and 3.10 in [3], the coercivity condition (4.9), together with the constraint qualifications (H1) and (H2), ensures that there exists $\rho > 0$ such that for each $\delta \in \mathcal{B}_\rho^{X^*\times Y}(0)$ there is a unique stationary point $(x^D_\delta, u^D_\delta, p^D_\delta, \mu^D_\delta)$ of $(\text{AO})^D_\delta$, where (x^D_δ, u^D_δ) is the solution. Moreover, there is a constant $r > 0$ such that

$$\begin{aligned}&\|x^D_{\delta'} - x^D_{\delta''}\|_{1,2}, \|u^D_{\delta'} - u^D_{\delta''}\|_2, \|p^D_{\delta'} - p^D_{\delta''}\|_{1,2}, \|\mu^D_{\delta'} - \mu^D_{\delta''}\|_{1,2} \le r\|\delta' - \delta''\|_{X^*\times Y}\\ &\qquad \forall\ \delta', \delta'' \in \mathcal{B}_\rho^{X^*\times Y}(0).\end{aligned} \tag{5.6}$$

Constants ρ and r depend on χ and $\widetilde{\gamma}$ in (H2) and (4.9), respectively; however, they can be chosen independently of $D \in \mathcal{D}_\beta$.

Note that $(\widehat{x}, \widehat{u}, \widehat{p}, \widehat{\mu})$ is a stationary point of $(\text{AO})^D_0$. So (5.6) implies, in particular,

$$\|\mu^D_\delta - \widehat{\mu}\|_{1,2} \le r\|\delta\|_{X^*\times Y}. \tag{5.7}$$

Set

$$\theta \in \left(0, \min\left\{\rho, \frac{\alpha}{4r}\sqrt{\frac{\beta^3}{6}}\right\}\right), \tag{5.8}$$

where $\beta > 0$ is given as in (4.8). Choose an arbitrary $\delta \in \mathcal{B}_\theta^{X^* \times Y}(0)$ and an arbitrary $D \in \mathcal{D}_\beta$. Denote

$$P_\delta^D := \{t \in [0,1] \mid \dot{\mu}_\delta^D(\cdot) = \text{const a.e. in a neighborhood of } t\}.$$

By Lemma 4.2 and (5.7)–(5.8), there exists $\beta' \in (0, \beta/2)$ such that

$$\text{meas } ([t', t''] \cap N_\alpha) \le \beta' < \beta/2 \quad \text{for any } [t', t''] \subset P_\delta^D. \tag{5.9}$$

Thus condition (4.8) is satisfied with a margin.

Let $\{\zeta_\delta^D := (x_\delta^D, u_\delta^D)\}$ be a sequence of solutions to $(\mathrm{AO})_\delta^D$ minimizing $J(x_\delta^D, u_\delta^D; \delta)$ with respect to $D \in \mathcal{D}_\beta$, i.e.,

$$\lim_D J(x_\delta^D, u_\delta^D; \delta) = \inf_{D \in \mathcal{D}_\beta} J(x_\delta^D, u_\delta^D; \delta) := \bar{J}_\delta. \tag{5.10}$$

It follows from (4.9) and (5.6) that $\bar{J}_\delta$ is finite and the set $\{(x_\delta^D, u_\delta^D) \in X^2 \mid D \in \mathcal{D}_\beta\}$ is weakly compact in X^2. Hence we can extract a weakly convergent subsequence, still denoted by $\{(x_\delta^D, u_\delta^D)\}$. Thus, there exists an element $\eta_\delta := (y_\delta, v_\delta) \in X^2$ such that

$$\begin{cases} u_\delta^D \rightharpoonup v_\delta & \text{weakly in } L^2(0,1;\mathbb{R}^m), \\ x_\delta^D \rightharpoonup y_\delta & \text{weakly in } W_0^{1,2}(0,1;\mathbb{R}^n), \text{ i.e., strongly in } C(0,1;\mathbb{R}^n). \end{cases} \tag{5.11}$$

Clearly, (y_δ, v_δ) satisfies the state equation. Moreover, in view of the strong convergence $x_\delta^D \to y_\delta$ in $C(0,1;\mathbb{R}^n)$, $D_x\widehat{\vartheta}(t)y_\delta(t) - (\widehat{\delta}_4(t) + \delta_4(t)) \le 0$ for all $t \in [0,1]$ and the set

$$D_\delta := \{t \in [0,1] \mid D_x\widehat{\vartheta}(t)y_\delta(t) - (\widehat{\delta}_4(t) + \delta_4(t)) < 0\} \tag{5.12}$$

satisfies condition (5.9). Hence (y_δ, v_δ) is feasible for $(\mathrm{AO})_\delta^{D_\delta}$ and $D_\delta \in \mathcal{D}_\beta$. We will show that (y_δ, v_δ) is the solution of $(\mathrm{AO})_\delta^{D_\delta}$. Indeed, taking $(x_\delta^D, v_\delta^D) \in \{(x_\delta^D, v_\delta^D)\}$, passing to the limit, and using (5.10) and (5.11), as well as (5.4) and (H3), we get

$$\begin{aligned} \bar{J}_\delta - J(y_\delta, v_\delta; \delta) &= \lim_D \left\{J(x_\delta^D, u_\delta^D; \delta) - J(y_\delta, v_\delta; \delta)\right\} \\ &= \lim_D \left\{ \frac{1}{2}((x_\delta^D, u_\delta^D), D^2\widehat{\mathcal{L}}\,(x_\delta^D, u_\delta^D)) - \frac{1}{2}((y_\delta, v_\delta), D^2\widehat{\mathcal{L}}\,(y_\delta, v_\delta)) \right. \\ &\qquad \left. - (\widehat{\delta}_1 + \delta_1, x_\delta^D - y_\delta) + (\widehat{\delta}_2 + \delta_2, u_\delta^D - v_\delta) \right\} \\ &= \lim_D \frac{1}{2}\left((u_\delta^D, D_{uu}^2\widehat{\mathcal{L}}\,u_\delta^D) - (v_\delta, D_{uu}^2\widehat{\mathcal{L}}\,v_\delta)\right) \\ &= \lim_D \frac{1}{2}\left((u_\delta^D - v_\delta), D_{uu}^2\widehat{\mathcal{L}}\,(u_\delta^D - v_\delta)\right) \ge 0. \end{aligned} \tag{5.13}$$

Since (y_δ, v_δ) is feasible for $(\mathrm{AO})_\delta^{D_\delta}$, (5.10) together with (5.13) implies

$$J(y_\delta, v_\delta; \delta) = \bar{J}_\delta, \tag{5.14}$$

which shows that (y_δ, v_δ) is the solution of $(\mathrm{AO})_\delta^{D_\delta}$.

Denote by $(q_\delta, \nu_\delta) \in Y^2$ the unique Lagrange multiplier of $(\mathrm{AO})_\delta^{D_\delta}$ associated with (y_δ, v_δ). We will show that $(y_\delta, v_\delta, q_\delta, \nu_\delta)$ is a stationary point of $(\mathrm{AO})_\delta$. The stationarity conditions for $(\mathrm{AO})_\delta^{D_\delta}$ take the form

$$
\left\{
\begin{array}{l}
\dot{q}_\delta(t) + D_x\widehat{f}^*(t)q_\delta(t) + D_xJ(y_\delta, v_\delta; \delta)(t) \\
\qquad +(D_x\widehat{f}^*(t)D_x\widehat{\vartheta}^*(t) + D_{xx}^2\widehat{\vartheta}(t)\widehat{f}(t) + D_{hx}^2\widehat{\vartheta}(t)\dot{\widehat{h}})\dot{\nu}_\delta(t) = 0, \\
q_\delta(1) = 0, \\
D_uJ(y_\delta, v_\delta; \delta)(t) + D_u\widehat{f}^*(t)q_\delta(t) + D_u\widehat{f}^*(t)D_x\widehat{\vartheta}^*(t)\dot{\nu}_\delta(t) = 0, \\
\nu_\delta(0)(D_x\widehat{\vartheta}(0)y_\delta(0) - (\widehat{\delta}_4(0) + \delta_4(0))) \\
\qquad + \displaystyle\int_0^1 \dot{\nu}_\delta(t)\frac{d}{dt}\left(D_x\widehat{\vartheta}(t)y_\delta(t) - (\widehat{\delta}_4(t) + \delta_4(t))\right)dt = 0,
\end{array}
\right. \tag{5.15}
$$

where $\nu_\delta \in K_{D_\delta}^+$, with

$$
K_{D_\delta}^+ = \left\{ \mu \in W^{1,2}(0,1;\mathbb{R}) \;\middle|\; \left\{
\begin{array}{l}
\mu(0) - \dot{\mu}(0) \geq 0,\ \dot{\mu}(\cdot) \text{ is nonincreasing on} \\
\quad \text{each subinterval of } [0,1] \setminus (N_\alpha \setminus D_\delta); \\
\text{if } 1 \notin \overline{N_\alpha \setminus D_\delta}, \text{ then there is a} \\
\quad \text{subinterval } (\tau, 1) \text{ such that } \dot{\mu}(t) \geq 0 \text{ for } t \in (\tau, 1).
\end{array}
\right. \right\} \tag{5.16}
$$

It follows from (5.15) that to prove that $(y_\delta, v_\delta, q_\delta, \nu_\delta)$ is a stationary point of $(\mathrm{AO})_\delta$, it is enough to show that $\nu_\delta \in K^+$, where K^+ is defined as in (3.9). Hence, we have to show that the sign and monotonicity conditions are satisfied by $\dot{\nu}_\delta(\cdot)$ on $(0,1)$. Choose any $\tau \in (0,1)$. In view of (5.9), there is a neighborhood $\mathcal{O}(\tau) \subset (0,1)$ of τ such that there exists $D(\tau) \in \mathcal{D}_\beta$ with the property that $D_\delta \cup \mathcal{O}(\tau) \subset D(\tau)$, where D_δ is defined as in (5.12). Consider problem $(\mathrm{AO})_\delta^{D(\tau)}$. It has a unique stationary point $(x_\delta^{D(\tau)}, u_\delta^{D(\tau)}, p_\delta^{D(\tau)}, \mu_\delta^{D(\tau)})$, where $(x_\delta^{D(\tau)}, u_\delta^{D(\tau)})$ is the solution. Note that (y_δ, v_δ) is feasible for $(\mathrm{AO})_\delta^{D(\tau)}$. Hence, by (5.14) it is the solution of that problem, i.e., $(x_\delta^{D(\tau)}, u_\delta^{D(\tau)}) = (y_\delta, v_\delta)$. Since, for a fixed $(x_\delta^{D(\tau)}, u_\delta^{D(\tau)})$, the element $(p_\delta^{D(\tau)}, \mu_\delta^{D(\tau)})$ satisfying (5.15) is unique, we find that $(q_\delta, \nu_\delta) = (p_\delta^{D(\tau)}, \mu_\delta^{D(\tau)})$ is the unique multiplier of $(\mathrm{AO})_\delta^{D(\tau)}$. In view of (5.16), it shows that $\dot{\nu}_\delta(\cdot)$ must be nonincreasing on $\mathcal{O}(\tau)$. Since $\tau \in (0,1)$ is arbitrary, $\dot{\nu}_\delta(\cdot)$ must be nonincreasing on $(0,1)$. Suppose now that $\dot{\nu}_\delta(t)$ is negative at some $t \in (0,1)$; then, by the monotonicity of $\dot{\nu}(\cdot)$, it must be negative on a subinterval $\mathcal{O}$ left of the final point 1, and we must have $\underline{D_x\widehat{\vartheta}(t)y_\delta(t) - (\widehat{\delta}_4(t) + \delta_4(t))} = 0$ on $\mathcal{O}$. There are two possible situations: either $1 \notin \overline{N_\alpha \setminus D_\delta}$ or $1 \in \overline{N_\alpha \setminus D_\delta}$. In the first case, (5.16) implies that there exists a subinterval $(\tau, 1)$ such that $\dot{\nu}_\delta(t) \geq 0$ for $t \in (\tau, 1)$. In the second case, by (5.9) there exists $\tau < 1$ such that $D(\tau) := (D_\delta \cup (\tau, 1)) \in \mathcal{D}_\beta$. Hence (y_δ, v_δ) is the solution of $(\mathrm{AO})_\delta^{D(\tau)}$, and by (5.16) we must have again $\dot{\nu}_\delta(t) \geq 0$ for $t \in [\tau, 1]$. This contradicts the assumption that $\dot{\nu}_\delta(t) < 0$ for some $t \in (0,1)$ and proves that $(y_\delta, v_\delta, q_\delta, \nu_\delta)$ is a stationary point of $(AO)_\delta$. Finally, since condition (4.9) is satisfied for $D = D_\delta$, Lemma 3.7 implies that (y_δ, v_δ) is a solution of $(\mathrm{AO})_\delta$.

Clearly, (5.6) implies (4.12). To show (4.13), note that it follows from (4.12) and (3.14) that, for $\theta > 0$ sufficiently small, we have $\{t \in [0,1] \mid \widehat{\vartheta}(t)y_\delta(t) - (\widehat{\delta}_4(t) + \delta_4(t)) = 0\} \subset M_\alpha$. Hence, using (H2) and (H3), as well as the regularity (4.6) of δ, we can repeat the proof of Proposition 6.6 in [8], and we arrive at (4.13). In view of Lemma 3.5, $\varsigma > \widehat{\varsigma}$. ∎

Proof of Proposition 4.4. Let $\delta \in \mathcal{B}^{\Delta}_{\theta}(0)$, where θ is given as in Lemma 4.3, and let $(y_\delta, v_\delta, q_\delta, \nu_\delta)$ be a corresponding stationary point of $(\mathrm{AO})_\delta$, satisfying (4.12). In view of (4.13), we have $\nu_\delta \in W^{2,2}_0(0,1;\mathbb{R})$. Hence, in a similar way as in (3.17) and (3.18), we can define

$$N^{\delta}_{\alpha/2} = [0,1] \setminus \overline{\{t \in [0,1] \mid -\ddot{\nu}_\delta \le \tfrac{\alpha}{2}\}},$$

$$\mathcal{E}^{\delta}_{\alpha/2} = \left\{ (y,v) \in X^2 \,\middle|\, \begin{cases} \dot{y}(t) - D_x\widehat{f}(t)y(t) - D_u\widehat{f}(t)v(t) = 0, \\ \langle D_x\widehat{\vartheta}(t), y(t)\rangle = 0 \ \forall t \in N^{\delta}_{\alpha/2}, \\ \langle D_x\widehat{\vartheta}(1), y(1)\rangle = 0 \ \text{ if } \dot{\widehat{\mu}}(1) > 0. \end{cases} \right\}$$

Using (5.7) and the same argument as in the proof of Lemma 4.2, we find that, for any subinterval $(t', t'') \subset N_\alpha \setminus N^{\delta}_{\alpha/2}$, the following estimate holds:

$$|t'' - t'| \le \left(12 \left(\frac{\|\nu_\delta - \widehat{\mu}\|_{1,2}}{\alpha/2} \right)^2 \right)^{1/3} \le \left(48 \left(\frac{r}{\alpha} \right)^2 \|\delta\|^2_{X^* \times Y} \right)^{1/3}.$$

If necessary, shrink $\theta > 0$, so that

$$\theta \le \left(\frac{1}{6} \right)^{1/2} \left(\frac{\beta}{4} \right)^{3/2} \frac{\alpha}{r},$$

where β is given in (4.8). Hence we get

$$|t'' - t'| \le \frac{\beta}{2} \quad \text{for any } (t', t'') \subset N_\alpha \setminus N^{\delta}_{\alpha/2} \text{ and any } \delta \in \mathcal{B}^{\Delta}_{\theta}(0).$$

Therefore, by Lemma 4.1 there exists a constant $\widetilde{\gamma} > 0$ such that

$$\begin{aligned} \Big((y,v), D^2\widehat{\mathcal{L}}\,(y,v) \Big) &\ge \widetilde{\gamma}(\|y\|^2_{1,2} + \|v\|^2_2) \\ &\forall\ (y,v) \in \mathcal{E}^{\delta}_{\alpha/2} \text{ and all } \delta \in \mathcal{B}^{\Delta}_{\theta}(0). \end{aligned} \tag{5.17}$$

Using (5.17) and repeating the argument of the proof of (4.12), but around δ rather than around 0, we find that there exists $\theta(\delta) > 0$, as well as $l > 0$ independent of δ, such that

$$\begin{aligned} &\|y_\phi - y_\delta\|_{1,2},\ \|v_\phi - v_\delta\|_2,\ \|q_\phi - q_\delta\|_{1,2},\ \|\nu_\phi - \nu_\delta\|_{1,2} \\ &\qquad \le l\ \|\phi - \delta\|_{X^* \times Y} \quad \forall\ \phi \in \mathcal{B}^{\Delta}_{\theta(\delta)}(\delta). \end{aligned} \tag{5.18}$$

Let us choose any $\delta', \delta'' \in \mathcal{B}^{\Delta}_{\theta}(0)$. For each $\delta_\sigma := (1-\sigma)\delta' + \sigma\delta''$, $\sigma \in [0,1]$, there exists $\theta(\delta_\sigma)$ such that (5.18) holds. The family $\{\mathcal{B}^{\Delta}_{\theta(\delta_\sigma)}(\delta_\sigma) \mid \sigma \in [0,1]\}$ constitutes a cover of the compact set $\{\delta_\sigma \in \Delta \mid \sigma \in [0,1]\}$. From this cover we can extract a finite cover $\{\mathcal{B}^{\Delta}_{\theta(\delta_{\sigma_i})}(\delta_{\sigma_i}) \mid \sigma = 1, \dots, q\}$, where $\delta_{\sigma_1} = \delta'$ and $\delta_{\sigma_q} = \delta''$. For any $1 \le j < q$, we can choose $\bar{\sigma}_j \in (\sigma_j, \sigma_{j+1})$ such that $\delta_{\bar{\sigma}_j} \in \mathcal{B}^{\Delta}_{\theta(\delta_{r_\sigma})}(\delta_{\sigma_j}) \cap \mathcal{B}^{\Delta}_{\theta(\delta_{\sigma_{j+1}})}(\delta_{\sigma_j+1})$. From (5.18), we get

$$\begin{aligned} &\|y_{\delta_{\bar{\sigma}_j}} - y_{\delta_{\sigma_j}}\|_{1,2},\ \|v_{\delta_{\bar{\sigma}_j}} - v_{\delta_{\sigma_j}}\|_2,\ \|q_{\delta_{\bar{\sigma}_j}} - q_{\delta_{\sigma_j}}\|_{1,2},\ \|\nu_{\delta_{\bar{\sigma}_j}} - \nu_{\delta_{\sigma_j}}\|_{1,2} \\ &\qquad \le l\ \|\delta_{\bar{\sigma}_j} - \delta_{\sigma_j}\|_{X^* \times Y} = (\bar{\sigma}_j - \sigma_j)\ l\ \|\delta'' - \delta'\|_{X^* \times Y}, \\ &\|y_{\delta_{\sigma_{j+1}}} - y_{\delta_{\bar{\sigma}_j}}\|_{1,2},\ \|v_{\delta_{\sigma_{j+1}}} - v_{\delta_{\bar{\sigma}_j}}\|_2,\ \|q_{\delta_{\sigma_{j+1}}} - q_{\delta_{\bar{\sigma}_j}}\|_{1,2},\ \|\nu_{\delta_{\sigma_{j+1}}} - \nu_{\delta_{\bar{\sigma}_j}}\|_{1,2} \\ &\qquad \le\ (\sigma_{j+1} - \bar{\sigma}_j)\ l\ \|\delta'' - \delta'\|_{X^* \times Y}. \end{aligned}$$

Thus

$$\begin{aligned}&\|y_{\delta_{\sigma_{j+1}}} - y_{\delta_{\sigma_j}}\|_{1,2},\ \|v_{\delta_{\sigma_{j+1}}} - v_{\delta_{\sigma_j}}\|_2,\ \|q_{\delta_{\sigma_{j+1}}} - q_{\delta_{\sigma_j}}\|_{1,2},\ \|\nu_{\delta_{\sigma_{j+1}}} - \mu_{\delta_{\sigma_j}}\|_{1,2}\\ &\qquad \le\ (\sigma_{j+1} - \sigma_j)\ l\ \|\delta'' - \delta'\|_{X^*\times Y}.\end{aligned}$$

Summing this inequality over $j = 1, \dots, q-1$, we arrive at (4.14). In particular, it implies that, for each $\delta \in \mathcal{B}^{\Delta}_{\theta}(0)$, the stationary point $(y_\delta, v_\delta, q_\delta, \nu_\delta)$ of $(\mathrm{AO})_\delta$ is unique in $\mathcal{B}^{X^2\times Y^2}_{\tau}(\widehat{x}, \widehat{u}, \widehat{p}, \widehat{\mu})$, where $\tau = l\,\theta$. □

Proof of Lemma 4.5. We have to show that each component $m_i(\xi, \lambda, h)$, $i = 1, \dots, 4$, of the vector function $m(\xi, \lambda, h)$ satisfies conditions (4.6). Let us confine ourselves to checking the needed properties of $m_4(\xi, \lambda, h)$. For the remaining components the estimates are similar. From (4.15) we get

$$m_4(\xi, \lambda, h) = \int_0^1 \Big[D_x\vartheta(\widehat{x}, \widehat{h}) - D_x\vartheta(x_\beta, \widehat{h})\Big]\, d\beta\ (x - \widehat{x}) - \int_0^1 D_h\vartheta(x, h_\beta) d\beta\ (h - \widehat{h}).$$

Hence we have

$$\begin{aligned}\frac{d^2}{dt^2} m_4(\xi, \lambda, h)(t) \quad &= \int_0^1 \frac{d^2}{dt^2}\Big[D_x\vartheta(\widehat{x}(t), \widehat{h}(t)) - D_x\vartheta(x_\beta(t), \widehat{h}(t))\Big]\, d\beta\ (x(t) - \widehat{x}(t))\\ &+2\int_0^1 \frac{d}{dt}\Big[D_x\vartheta(\widehat{x}(t), \widehat{h}(t)) - D_x\vartheta(x_\beta(t), \widehat{h}(t))\Big]\, d\beta\ (\dot{x}(t) - \dot{\widehat{x}}(t))\\ &+\int_0^1 \Big[D_x\vartheta(\widehat{x}(t), \widehat{h}(t)) - D_x\vartheta(x_\beta(t), \widehat{h}(t))\Big]\, d\beta\ (\ddot{x}(t) - \ddot{\widehat{x}}(t))\\ &-\int_0^1 \frac{d^2}{dt^2} D_h\vartheta(x(t), h_\beta(t)) d\beta\ (h(t) - \widehat{h}(t))\\ &-2\int_0^1 \frac{d}{dt} D_h\vartheta(x(t), h_\beta(t)) d\beta\ (\dot{h}(t) - \dot{\widehat{h}}(t))\\ &-\int_0^1 D_h\vartheta(x(t), h_\beta(t)) d\beta\ (\ddot{h}(t) - \ddot{\widehat{h}}(t)).\end{aligned} \tag{5.19}$$

It can be easily seen that, in view of (I), the first three components on the right-hand side of (5.19) tend to zero as $\|x - \widehat{x}\|_{1,2} \to 0$, provided that condition (4.2) is satisfied. Similarly, it follows from (3.1) and (4.2) that the remaining three terms tend to zero as $\|h - \widehat{h}\|_H \to 0$. Thus, choosing $\tau > 0$ and $\pi' > 0$ sufficiently small, we get $\|d^2/dt^2\ m_4(\xi, \lambda, h)\|_\infty \le s$. Using the same argument, we find that the remaining components of $m(\xi, \lambda, h)$ also satisfy the required regularity conditions. Hence $m(\xi, \lambda, h)$ belongs to Δ and condition (B6) is satisfied. □

Proof of Corollary 4.8. Let us set $\varpi = \min\{\pi, \frac{\alpha}{\ell}\sqrt{\frac{\beta^3}{12}}\}$. By Theorem 4.6, for any $h \in \mathcal{B}^H_\varpi(\widehat{h})$, there is a unique stationary point (x_h, u_h, p_h, μ_h), whereas by (4.16) and Lemma 4.2 the sets $D_h := \{t \in [0,1] \mid \vartheta(x_h(t), h(t)) < 0\}$ satisfy condition (4.8). Hence, by Lemma 4.1 we get

$$\begin{aligned}&\Big((z, w), D^2\widehat{\mathcal{L}}\,(z, w)\Big) \ge \widetilde{\gamma}(\|z\|^2_{1,2} + \|w\|^2_2)\\ &\qquad \forall\ (z, w) \in \mathcal{E}_{\alpha, D_h} \text{ and all } h \in \mathcal{B}^H_\varpi(\widehat{h}).\end{aligned} \tag{5.20}$$

Denote

$$D^2\mathcal{L}^h = D^2_{(x,u)(x,u)}\mathcal{L}(x_h,u_h,p_h,\mu_h;h),$$

$$\mathcal{E}^h_\alpha = \left\{ (y,v)\in X^2 \,\middle|\, \left\{ \begin{array}{l} \dot y(t) - D_x f(x_h,u_h,h)(t)y(t) - D_u f(x_h,u_h,h)(t)v(t) = 0,\\ \langle D_x\vartheta(x_h,h)(t), y(t)\rangle = 0 \ \forall t\in N_\alpha\setminus D_h,\\ \langle D_x\vartheta(x_h,h)(t), y(t)\rangle = 0 \text{ if } \dot\mu_\delta(1)>0, \end{array}\right.\right\}$$

$$\mathcal{C}^h_\alpha \left[\begin{array}{c} y\\ v\\ \kappa\end{array}\right] = \left[\begin{array}{c} \dot y - D_x f(x_h,u_h,h)y - D_u f(x_h,u_h,h)v\\ D_x\vartheta(x_h,h)\,y + T_\alpha\kappa\end{array}\right].$$

We are going to show that, for $\varpi>0$ sufficiently small

$$\big((y,v), D^2\mathcal{L}^h(y,v)\big) \ge \frac{1}{2}\widetilde\gamma(\|y\|^2_{1,2}+\|v\|^2_2) \quad \forall\ (y,v)\in\mathcal{E}^h_\alpha, \tag{5.21}$$

which, by Lemma 3.7 shows that (x_h,u_h) is a second order local minimizer of $(\mathrm{O})_h$. Thus, the corollary will be proved.

By Lemma 3.3 the map

$$\mathcal{C}_\alpha\mathcal{C}^*_\alpha:\ L^2(0,1;\mathbb{R}^n)\times W^{1,2}(0,1;\mathbb{R}) \to L^2(0,1;\mathbb{R}^n)\times W^{1,2}(0,1;\mathbb{R})$$

is invertible. For any $(y,v)\in X^2$ define the following new variable:

$$\left[\begin{array}{c} z\\ w\\ \kappa\end{array}\right] = \Big[\mathcal{C}^*_\alpha\,(\mathcal{C}_\alpha\mathcal{C}^*_\alpha)^{-1}\,\mathcal{C}^h_\alpha + \big(I - \mathcal{C}^*_\alpha(\mathcal{C}_\alpha\mathcal{C}^*_\alpha)^{-1}\mathcal{C}_\alpha\big)\Big]\left[\begin{array}{c} y\\ v\\ 0\end{array}\right].$$

We get

$$\mathcal{C}_\alpha\left[\begin{array}{c} z\\ w\\ \kappa\end{array}\right] = \mathcal{C}^h_\alpha\left[\begin{array}{c} y\\ v\\ 0\end{array}\right].$$

Since in view of (3.15), $T_\alpha(t)=0$ for $t\in M_\alpha$, then for $(y,v)\in\mathcal{E}^h_\alpha$ we get

$$\begin{aligned} &\dot z - D_x\widehat f\, z - D_u\widehat f\, w = \dot y - D_x f(x_h,u_h,h)y - D_u f(x_h,u_h,h)v = 0,\\ &D_x\widehat\vartheta(t)z(t) = D_x\vartheta(x_h(t),h(t))y(t) = 0 \quad \text{for } t\in N_\alpha\setminus D_h. \end{aligned}$$

Moreover, it follows from (4.17) that, if $\dot{\widehat\mu}(1)>0$, then there exists $\varpi>0$ such that $\mu_h(1)>0$ for all $h\in\mathcal{B}^H_\varpi(\widehat h)$. That implies $\langle\vartheta(x_h(1),h(1)),z(1)\rangle = 0$. Thus, $(z,w)\in\mathcal{E}_{\alpha,D_h}$, i.e., it satisfies (5.20). On the other hand

$$\left[\begin{array}{c} \Delta y\\ \Delta v\\ -\kappa\end{array}\right] := \left[\begin{array}{c} y\\ v\\ 0\end{array}\right] - \left[\begin{array}{c} z\\ w\\ \kappa\end{array}\right] = \mathcal{C}^*_\alpha\,(\mathcal{C}_\alpha\mathcal{C}^*_\alpha)^{-1}\,\big(\mathcal{C}_\alpha - \mathcal{C}^h_\alpha\big)\left[\begin{array}{c} y\\ v\\ 0\end{array}\right]. \tag{5.22}$$

Using Young's inequality, we get

$$\begin{aligned} \big((y,v), D^2\mathcal{L}^h\ (y,v)\big) &= \Big((y,v),(D^2\mathcal{L}^h - D^2\widehat{\mathcal{L}})\ (y,v)\Big)\\ &+ \Big((z,w), D^2\widehat{\mathcal{L}}\ (z,w)\Big) + 2\Big((z,w), D^2\widehat{\mathcal{L}}\ (\Delta y,\Delta v)\Big) + \Big((\Delta y,\Delta v), D^2\widehat{\mathcal{L}}\ (\Delta y,\Delta v)\Big)\\ &\ge \frac{3}{4}\Big((z,w), D^2\widehat{\mathcal{L}}\ (z,w)\Big) - \Big|\Big((y,v),(D^2\mathcal{L}^h - D^2\widehat{\mathcal{L}})\ (y,v)\Big)\Big|\\ &-3\Big|\Big((\Delta y,\Delta v), D^2\widehat{\mathcal{L}}\ (\Delta y,\Delta v)\Big)\Big|. \end{aligned}$$

By (4.17) and (5.22) we have

$$\left|\Big((y,v),(D^2\mathcal{L}^h - D^2\widehat{\mathcal{L}})\ (y,v)\Big)\right| \to 0 \quad \text{and} \quad \|(\Delta y, \Delta v)\|_{X^2} \to 0,$$

uniformly in h, as $h \to 0$. Hence, in view of (5.20), condition (5.21) is satisfied, provided that ϖ is sufficiently small. ∎

Appendix A. Proof of Lemma 3.7. In the indirect proof, we follow the proof of Theorem 4.1 in [1]. Suppose that (3.19) is satisfied for $\alpha = 0$, but (3.20) is violated. Then, there exists a sequence $\{(x_j, u_j)\}$ of feasible pairs $(x_j, u_j) \in X^\infty$ such that $(x_j, u_j) \to (\widehat{x}, \widehat{u})$ and

$$F(x_j, u_j, \widehat{h}) \le F(\widehat{x}, \widehat{u}, \widehat{h}) + o(\|(x_j - \widehat{x}, u_j - \widehat{u})\|^2_{X^2}). \tag{A.1}$$

Denote $\epsilon_j := \|u_j - \widehat{u}\|_2$ and set $v_j = \epsilon_j^{-1}(u_j - \widehat{u})$. Thus, $\|v_j\|_2 = 1$. Let z_j be the solution of the linearized equation

$$\dot{z}_j(t) - D_x\widehat{f}(t)z_j(t) - D_u\widehat{f}(t)v_j(t) = 0, \quad z_j(0) = 0. \tag{A.2}$$

By the well-known result for differential equations we have

$$\|(x_j - \widehat{x})\|_{1,2} = O(\epsilon_j) \quad \text{and} \quad \|(x_j - \widehat{x}) - \epsilon_j z_j\|_{1,2} = o(\|x_j - \widehat{x}\|_{1,2}) = o(\epsilon_j). \tag{A.3}$$

From (A.1) and (A.3) we obtain

$$(D_xF(\widehat{x}, \widehat{u}, \widehat{h}), z_j) + (D_uF(\widehat{x}, \widehat{u}, \widehat{h}), v_j) \le O(\|(x_j - \widehat{x}, u_j - \widehat{u})\|_{X^2}) = O(\epsilon_j). \tag{A.4}$$

Since $\|v_j\|_2 = 1$, we can extract from $\{v_j\}$ a weakly convergent subsequence, still denoted by $\{v_j\}$. So, there is a pair $(\bar{y}, \bar{v})$ such that

$$\begin{cases} v_j \rightharpoonup \bar{v} & \text{weakly in } L^2(0,1,\mathbb{R}^m), \\ z_j \rightharpoonup \bar{z} & \text{weakly in } W^{1,2}(0,1,\mathbb{R}^n), \text{ i.e., strongly in } C(0,1;\mathbb{R}^n), \end{cases} \tag{A.5}$$

where $(\bar{z}, \bar{v})$ is a solution of (A.2). Passing to the limit in (A.4), we find that

$$(D_xF(\widehat{x}, \widehat{u}, \widehat{h}), \bar{z}) + (D_uF(\widehat{x}, \widehat{u}, \widehat{h}), \bar{v}) \le 0.$$

Note that, by stationarity of the Lagrangian we have

$$(D_xF(\widehat{x}, \widehat{u}, \widehat{h}), \bar{z}) + (D_uF(\widehat{x}, \widehat{u}, \widehat{h}), \bar{v}) = -\left(\dot{\widehat{\mu}}, \frac{d}{dt}(D_x\widehat{\vartheta}\bar{z})\right) = -\dot{\widehat{\mu}}(1)D_x\widehat{\vartheta}(1)\bar{z}(1) + (\ddot{\widehat{\mu}}, D_x\widehat{\vartheta}\bar{z}).$$

Hence

$$-\dot{\widehat{\mu}}(1)D_x\widehat{\vartheta}(1)\bar{z}(1) + (\ddot{\widehat{\mu}}, D_x\widehat{\vartheta}\bar{z}) \le 0. \tag{A.6}$$

From the definition (3.14) we have $\vartheta(x_j(t)) - \vartheta(\widehat{x}(t)) \le 0$ for all $t \in M_0$. Hence, by (A.3) and (A.5) we get $D_x\widehat{\vartheta}(t)\bar{z}(t) \le 0$ for all $t \in M_0$. Similarly, if $\dot{\widehat{\mu}}(1) > 0$, then $\vartheta(x_j(1)) - \vartheta(\widehat{x}(1)) \le 0$, and hence $D_x\widehat{\vartheta}(1)\bar{z}(1) \le 0$. Thus, in view of (3.17) and (3.18), (A.6) implies

$$D_x\widehat{\vartheta}(t)\bar{z}(t) = 0 \ \text{ for } t \in N_0 \quad \text{and} \quad D_x\widehat{\vartheta}(1)\bar{z}(1) = 0 \ \text{ if } \ \dot{\widehat{\mu}}(1) > 0,$$

which together with (A.2) shows that $(\bar{z}, \bar{v}) \in \mathcal{E}_0$.

Since $\widehat{\mu} \in K^+$, it follows from (3.7) and (A.3) that

$$\begin{aligned} & F(x_j, u_j, \widehat{h}) - F(\widehat{x}, \widehat{u}, \widehat{h}) \geq \mathcal{L}(x_j, u_j, \widehat{p}, \widehat{\mu}) - \mathcal{L}(\widehat{x}, \widehat{u}, \widehat{p}, \widehat{\mu}) \\ & = \frac{1}{2}\left((x_j - \widehat{x}, u_j - \widehat{u}), D^2\widehat{\mathcal{L}}\,(x_j - \widehat{x}, u_j - \widehat{u})\right) + o(\|(x_j - \widehat{x}, u_j - \widehat{u})\|^2_{X^2}) \\ & = \frac{1}{2}\epsilon_j^2\left(((z_j, v_j), D^2\widehat{\mathcal{L}}\,(z_j, v_j)) + O(\epsilon_j)\right). \end{aligned} \tag{A.7}$$

Substituting (A.1) into (A.7) and passing to the limit, we get

$$\liminf_{j\to\infty}\left((z_j, v_j), D^2\widehat{\mathcal{L}}\,(z_j, v_j)\right) \leq \limsup_{j\to\infty}\left((z_j, v_j), D^2\widehat{\mathcal{L}}\,(z_j, v_j)\right) \leq 0.$$

Since, in view of (5.4) and (H3), $((z, v), D^2\widehat{\mathcal{L}}\,(z, v))$ is weakly lower semicontinuous in X^2, by (A.5) we get

$$0 \geq \liminf_{j\to\infty}\left((z_j, v_j), D^2\widehat{\mathcal{L}}\,(z_j, v_j)\right) \geq \left((\bar{z}, \bar{v}), D^2\widehat{\mathcal{L}}\,(\bar{z}, \bar{v})\right).$$

Since $(\bar{z}, \bar{v}) \in \mathcal{E}_0$, (3.19) implies that

$$(\bar{z}, \bar{v}) = (0, 0). \tag{A.8}$$

Thus

$$\lim_{j\to\infty}\left((z_j, v_j), D^2\widehat{\mathcal{L}}\,(z_j, v_j)\right) = 0 = \left((\bar{z}, \bar{v}), D^2\widehat{\mathcal{L}}\,(\bar{z}, \bar{v})\right).$$

In particular we get $\lim_{j\to\infty}(v_j, D^2_{uu}\widehat{\mathcal{L}}\ v_j) = (\bar{v}, D^2_{uu}\widehat{\mathcal{L}}\ \bar{v})$. By (H3) and (A.5), it implies that $v_j \to \bar{v}$ strongly in $L^2(0, 1; \mathbb{R}^m)$. Since $\|v_j\|_2 = 1$, we get $\|\bar{v}\|_2 = 1$, which contradicts (A.8) and completes the proof. ∎

REFERENCES

[1] F. Bonnans and A. Hermant, *No-Gap Second-Order Optimality Conditions for Optimal Control Problems with a Single State Constraint and Control*, Research Report 5837, INRIA, Le Chesnay, 2006.

[2] A. L. Dontchev, *Implicit function theorems for generalized equations*, Math. Program., 70 (1995), pp. 91–106.

[3] A. L. Dontchev and W. W. Hager, *Lipschitzian stability for state constrained nonlinear optimal control*, SIAM J. Control Optim., 36 (1998), pp. 698–718.

[4] A. L. Dontchev and K. Malanowski, *A characterization of Lipschitzian stability in optimal control*, in Calculus of Variations and Optimal Control, A. Ioffe, S. Reich, and I. Shafrir, eds., Chapman Hall/CRC Res. Notes Math. 411, Chapman & Hall/CRC, Boca Raton, FL, 2000, pp. 62–76.

[5] W. W. Hager, *Lipschitz continuity for constrained processes*, SIAM J. Control Optim., 17 (1979), pp. 321–338.

[6] R. F. Hartl, S. P. Sethi, and R. G. Vickson, *A survey of the maximum principles for optimal control problems with state constraints*, SIAM Rev., 37 (1995), pp. 181–218.

[7] K. Malanowski, *Two-norm approach in stability and sensitivity analysis for optimization and optimal control problems*, Adv. Math. Sci. Appl., 2 (1993), pp. 397–443.

[8] K. Malanowski, *Stability and sensitivity of solutions to nonlinear optimal control problems*, Appl. Math. Optim., 32 (1995), pp. 111–141.

[9] K. Malanowski, *Stability and Sensitivity Analysis for Optimal Control Problems with Control–State Constraints*, Dissertationes Math. (Rozprawy Mat.) 394, Polska Akademia Nauk, Instytut Matematyczny, Warszawa, 2001.

[10] K. Malanowski, *On normality of Lagrange multipliers for state constrained optimal control problems*, Optimization, 52 (2003), pp. 75–91.
[11] K. Malanowski, *Sufficient optimality conditions in stability analysis for state-constrained optimal control*, Appl. Math. Optim., 55 (2007), pp. 255–271.
[12] K. Malanowski, *Stability and sensitivity analysis for linear-quadratic optimal control subject to state constraints*, Optimization, 56 (2007), pp. 463–478.
[13] H. Maurer, *First and second order sufficient optimality conditions in mathematical programming and optimal control*, Math. Programming Study, 14 (1981), pp. 163–177.
[14] L. W. Neustadt, *Optimization: A Theory of Necessary Conditions*, Princeton University Press, Princeton, NJ, 1976.
[15] J. V. Outrata and Z. Schindler, *An augmented Lagrangian method for a class of convex optimal control problems*, Prob. Control Inform. Theory, 10 (1980), pp. 67–81.
[16] S. M. Robinson, *Strongly regular generalized equations*, Math. Oper. Res., 5 (1980), pp. 43–62.

SIAM J. OPTIM.
Vol. 18, No. 3, pp. 946–960

ASPLUND DECOMPOSITION OF MONOTONE OPERATORS*

JONATHAN BORWEIN† AND HERRE WIERSMA‡

Abstract. We establish representations of a monotone mapping as the sum of a maximal subdifferential mapping and a "remainder" monotone mapping, where the remainder is "acyclic" in the sense that it contains no nontrivial subdifferential component. This is the nonlinear analogue of a skew linear operator. Examples of indecomposable and acyclic operators are given. In particular, we present an explicit nonlinear acyclic operator.

Key words. monotone operators, cyclic monotonicity, decompositions, convex subgradients, acyclic operators

AMS subject classifications. 47H04, 52A41

DOI. 10.1137/060658357

1. Introduction.

Let X be a Banach space and X^* its topological dual. We denote the closed unit ball in X by B_X or B. Recall that a *monotone operator* $T : X \rightrightarrows X^*$ is a mapping that satisfies

$$\langle x^* - y^*, x - y\rangle \geq 0$$

whenever $x^* \in T(x)$ and $y^* \in T(y)$.

The *domain* of T is $\operatorname{dom} T = \{x \in \mathbb{R}^n \mid T(x) \neq \emptyset\}$, and the *range* of T is $\operatorname{ran} T = \{x^* \in \mathbb{R}^n \mid x^* \in T(x) \text{ for some } x \in \operatorname{dom} T\}$. The *graph* of T is the set $\operatorname{gr} T := \{(x, x^*) \in \mathbb{R}^n \times \mathbb{R}^n \mid x^* \in T(x)\}$. Of particular interest are maximal monotone operators: T is said to be *maximal monotone* if $\operatorname{gr} T \subset \operatorname{gr} S$ with S monotone implies that $T = S$.

In general, T could be a multivalued mapping on an infinite-dimensional space; however, the phenomena we wish to discuss are poorly understood, even for single valued mappings in $\mathbb{R}^n$. We will restrict ourselves largely to this setting where T is single valued, and X and X^* both are $\mathbb{R}^n$; in the following, the notation $T : \operatorname{dom} T \subset X \to X^*$ (single arrow) always denotes a single valued operator. This is not an unreasonable restriction, since the results that hold in $\mathbb{R}^n$, such as continuity or differentiability theorems, usually have a reasonable extension at least to separable Asplund spaces [4]. Moreover, in $\mathbb{R}^n$, T is almost everywhere single valued on $\operatorname{int} \operatorname{dom} T$, from which most of our results naturally extend to the multivalued case. Further background and references may be found in [2, 3, 4, 5].

One important instance of a maximal monotone operator is the subdifferential of a convex function. Let f be a proper convex lower semicontinuous function on $\mathbb{R}^n$. Then the *subdifferential* $\partial f : \operatorname{dom} \partial f \subset \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is the monotone mapping

$$\partial f(x) = \{x^* \in \mathbb{R}^n \mid \langle x^*, y - x\rangle + f(x) \leq f(y) \text{ for all } y \in \mathbb{R}^n\}.$$

*Received by the editors April 27, 2006; accepted for publication (in revised form) April 4, 2007; published electronically October 4, 2007.
http://www.siam.org/journals/siopt/18-3/65835.html
†Faculty of Computer Science, Dalhousie University, 6050 University Avenue, Halifax, NS B3H 1W5, Canada (jborwein@cs.dal.ca).
‡Department of Mathematics and Statistics, Dalhousie University, 6050 University Avenue, Halifax, NS B3H 3J5, Canada (hwiersma@cs.dal.ca).

Subdifferential mappings enjoy a variety of nice properties: they are single valued on large sets and automatically maximal monotone [11, 4, 14] and seemingly belong to all classes of well-behaved maximal monotone operators in nonreflexive spaces (see [8, 9, 12, 13, 14]. Thus, it appears that if $T = \partial f + R$ possesses any pathology, it is contributed by R. For an arbitrary monotone mapping T, it is therefore appealing to consider decompositions of the form $T = \partial f + R$, where R is a "remainder" to be made as small as possible in some sense. This is an extension of the decomposition of a linear operator into its symmetric and skew parts: $L = (L + L^*)/2 + (L - L^*)/2$.

The "nicest" form for R to take is the zero mapping, in which case T is just a subdifferential map. Barring that, perhaps the next simplest form for R to take is a *skew* or *skew-like* mapping. We investigate in section 2 when such a decomposition is possible. Examples of operators for which this decomposition is not possible are given in section 3. Even if R does not take such a simple form, a modernized version of a 1970 result of Asplund (see [1, 4]) shows that we can find a decomposition with R "acyclic," as described in section 4. Little is known about the properties of such acyclic mappings, however. We give the first explicit example of a nonlinear acyclic operator $\widehat{S} : \mathbb{R}^2 \to \mathbb{R}^2$ in section 5. We further explore this mapping in section 6 and conclude with some open questions.

Since our central goal is to better understand acyclicity, little will be lost if the reader assumes throughout that every monotone operator is everywhere defined and single valued.

2. Skew decompositions. In this section we introduce various weakenings of the notion of a skew symmetric linear mapping and then link them to the properties of an associated function f_T due to Fitzpatrick as defined later in this paper. A mapping $SL : \operatorname{dom} SL \subset \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is said to be *skew-like* if $\langle x^*, x\rangle = 0$ for all $(x, x^*) \in \operatorname{gr} SL$, and $S : \operatorname{dom} S \subset \mathbb{R}^n \to \mathbb{R}^n$ is *skew* if it is linear and $\langle Sx, x\rangle = 0$ for all $x \in \operatorname{dom} S$. We allow that $\operatorname{dom} S \neq \mathbb{R}^n$; in this case we require that $S = \widehat{S}|_{\operatorname{dom} S}$ for some skew linear $\widehat{S} : \mathbb{R}^n \to \mathbb{R}^n$. Thus, skew mappings are ab initio restrictions of skew and linear mappings.

FACT 1. *Let* $0 \in \operatorname{int} \operatorname{dom} S$.

(1) *If* $S : \operatorname{dom} S \subset \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ *is monotone and skew-like, then it is skew linear on* $\operatorname{dom} S$.

(2) *If* $S : \operatorname{dom} S \subset \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ *is monotone, and* $-S$ *is monotone with* $0 \in S(0)$, *then* S *is skew linear on* $\operatorname{dom} S$.

Proof. (1) Using monotonicity and the fact that $\langle x, x^*\rangle = 0$ when $x^* \in S(x)$, we have $\langle x^*, y\rangle \leq -\langle y^*, x\rangle$ for all $(x, x^*), (y, y^*) \in \operatorname{gr} S$.

Choose $\varepsilon > 0$ so that $\varepsilon B \subset \operatorname{int} \operatorname{dom} S$, where B is the closed unit ball in $\mathbb{R}^n$. For $y, z \in \varepsilon B$ choose $y_1^* \in S(y)$, $y_2^* \in S(-y)$, and $z^* \in S(z)$. Then $\langle y_1^*, z\rangle \leq -\langle z^*, y\rangle$ and $\langle z^*, -y\rangle \leq -\langle y_2^*, z\rangle$, which combine to give

$$\langle y_1^* + y_2^*, z\rangle \leq 0 \qquad \text{for all } z \in \varepsilon B.$$

Hence $y_1^* = -y_2^*$ for all $y_1^* \in S(y)$ and $y_2^* \in S(-y)$, so $S(y)$ is singleton with $S(y) = -S(-y)$ whenever $y \in \varepsilon B$.

Let $(x, x^*) \in \operatorname{gr} S$, $y \in \varepsilon B$. Then

$$\langle x^*, y\rangle \leq -\langle S(y), x\rangle = \langle S(-y), x\rangle \leq -\langle x^*, -y\rangle = \langle x^*, y\rangle,$$

so $\langle x^*, y\rangle = -\langle S(y), x\rangle$. Suppose $(x_1, x_1^*), (x_2, x_2^*), (\alpha x_1 + \beta x_2, w^*) \in \operatorname{gr} S$. Then

$$\begin{aligned}\langle w^*, y\rangle &= -\langle S(y), \alpha x_1 + \beta x_2\rangle = -\alpha\langle S(y), x_1\rangle - \beta\langle S(y), x_2\rangle \\ &= \alpha\langle x_1^*, y\rangle + \beta\langle x_2^*, y\rangle = \langle \alpha x_1^* + \beta x_2^*, y\rangle\end{aligned}$$

for all $y \in \varepsilon B$, so that $w^* = \alpha x_1^* + \beta x_2^*$. Choosing $x_2 = x_1$ and $\alpha + \beta = 1$ shows that S is single valued on $\operatorname{dom} S$. That is, $S(\alpha x + \beta y) = \alpha S(x) + \beta S(y)$ whenever $x, y, \alpha x + \beta y \in \operatorname{dom} S$.

Since $\varepsilon B \subset \operatorname{dom} S$, it is clear that there is a unique skew linear extension $\widehat{S}$ of S to the whole space: $\widehat{S}(x) = (\|x\|/\varepsilon) S(\varepsilon x/\|x\|)$.

(2) If $x^* \in S(x)$, then

$$\langle x^*, x\rangle = \langle x^* - 0, x - 0\rangle = 0$$

since $0 \in S(0)$ and both S and $-S$ are monotone. So S is skew-like and monotone, and we can apply (1) to see that S is skew linear on $\operatorname{dom} S$. □

We will say that a monotone operator $T : \operatorname{dom} T \subset \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is *weakly decomposable* if it can be written as the sum of a (possibly zero) skew-like operator and the subgradient of a proper lower semicontinuous convex function: $T = S + \partial f$; and *decomposable* if the skew-like part is actually skew. If T is not decomposable, we say that it is *indecomposable*. For example, the addition of a skew mapping to the subgradient of any norm produces a multivalued decomposable maximal monotone mapping.

Note that a skew-like operator need not be monotone. Note also that if $T(x)$ is single valued and nonempty, so necessarily is $S(x)$ and $\partial f(x)$.

For the following, we use the notation $DT(x)$ for the Jacobian matrix of $T : \mathbb{R}^n \to \mathbb{R}^n$ at x, and we note that T is $\mathcal{C}^1$ (Fréchet or, equivalently in finite dimensions, Gâteaux) on an open set C if and only if the mapping $x \to DT(x)$ is continuous on C.

FACT 2. *Let $T : \operatorname{dom} T \subset \mathbb{R}^n \to \mathbb{R}^n$ be a continuously Fréchet differentiable maximal monotone mapping on an open domain. Then the decomposition $T = S + \nabla f$ into a skew component S and a subdifferential component $\partial f = \{\nabla f\}$ is unique when it exists.*

Proof. From now on we will identify $\{\nabla f\}$ and ∇f. Suppose $T = S + \nabla f = S_1 + \nabla g$. Then $S(x) - S_1(x) = \nabla g(x) - \nabla f(x)$. Differentiating gives

$$S - S_1 = \nabla^2(g - f)(x);$$

the left-hand side is a skew matrix, and the right-hand side is symmetric, so both must be zero matrices. □

A useful observation is the following.

FACT 3. *Let $T : \operatorname{dom} T \subset \mathbb{R}^n \to \mathbb{R}^n$ be continuously Fréchet differentiable on an open convex set $C \subset \operatorname{dom} T$ with $0 \in C$ and $T(0) = 0$. Then T is monotone (resp., skew) on C if and only if $DT(z)$ is positive semidefinite (resp., skew) throughout C.*

Proof. We prove only the skew case; the monotone case is similar. Let $DT(z)$ be skew for each z in the interior of C, and take $x, y \in C \subset \operatorname{dom} T$. The mean-value theorem then provides $z \in [x, y]$ with

$$\langle T(x) - T(y), x - y\rangle = \langle DT(z)(x - y), x - y\rangle = 0,$$

so T and $-T$ are monotones. Fact 1 shows that T is skew linear. On the other hand, suppose T is skew, with $x \in \operatorname{int} \operatorname{dom} T$. Fixing h, we see that

$$\langle th, DT(x + sh)\ th\rangle = \langle T(x + th) - T(x), th\rangle = 0$$

for some $0 < s < t$. Thus

$$\langle h, DT(x + sh)\ h\rangle = 0;$$

letting $t \to 0$ shows that $DT(x)$ is skew. ∎

Define *Fitzpatrick's last function* f_T relative to a point $a \in \operatorname{int}\operatorname{dom} T$ by

$$f_T(x; a) := \int_0^1 \langle T\big(a + t(x - a)\big), x - a\rangle\, dt.$$

(This construction was suggested to the authors by Simon Fitzpatrick just months before his death in 2004.) We use the notation $f_T(x) := f_T(x; 0)$, where $0 \in \operatorname{int}\operatorname{dom} T$.

We may use f_T to characterize both weak decomposability and decomposability. We start with a technical lemma.

LEMMA 1. *For any continuously Fréchet differentiable monotone operator* $T : \operatorname{dom} T \subset \mathbb{R}^n \to \mathbb{R}^n$ *with* $0 \in \operatorname{int}\operatorname{dom} T$*, it is always the case that* $S := T - \nabla f_T$ *is skew-like on* $\operatorname{int}\operatorname{dom} T$*.*

Proof. Fix $x, y \in \operatorname{int}\operatorname{dom} T$, and define

$$h(t) := \langle T(tx), ty\rangle.$$

We check that

$$\langle T(x), y\rangle = h(1) - h(0) = \int_0^1 t\,\langle DT(tx)x, y\rangle\, dt + \int_0^1 \langle T(tx), y\rangle\, dt \tag{2.1}$$

and

$$\langle \nabla f_T(x), y\rangle = \int_0^1 t\,\langle DT(tx)^T x, y\,\rangle\, dt + \int_0^1 \langle T(tx), y\rangle\, dt; \tag{2.2}$$

we can switch the order of integration and differentiation since $(x, t) \to \langle T(tx), x\rangle$ is continuous. Then $S := T - \nabla f_T$ is skew-like, since $\langle T(x), x\rangle = \langle \nabla f_T(x), x\rangle$. ∎

Throughout the rest of this section we assume $\operatorname{dom} T$ is open so as to avoid technical complications at boundary points.

THEOREM 2 (weak decomposability). *Suppose* T *is a continuously Fréchet differentiable maximal monotone operator* $T : \operatorname{dom} T \subset \mathbb{R}^n \to \mathbb{R}^n$ *for which* $0 \in \operatorname{int}\operatorname{dom} T = \operatorname{dom} T$*. Then the following are equivalent:*

(1) T *is weakly decomposable on* $\operatorname{dom} T$*,*

(2) f_T *is convex on* $\operatorname{dom} T$*.*

Proof. Letting $S := T - \nabla f_T$, Lemma 1 shows that S is skew-like. Hence if f_T is convex, T is weakly decomposable.

Conversely, suppose that $T = \nabla g + S$ with g convex and S skew-like. Then $f_{\nabla g} = f_T$ as is seen by writing $h(1) - h(0) = \int_0^1 h'(t)\, dt$ with $h := t \mapsto g(xt)$, which implies that $g - g(0) = f_T$ and we are done. ∎

THEOREM 3 (decomposability). *Suppose we have a continuously differentiable maximal monotone operator* $T : \operatorname{dom} T \subset \mathbb{R}^n \to \mathbb{R}^n$ *for which* $0 \in \operatorname{int}\operatorname{dom} T = \operatorname{dom} T$*. Then* T *is decomposable on* $\operatorname{dom} T$ *if and only if* $T - \nabla f_T$ *is skew on* $\operatorname{dom} T$*.*

Proof. Without loss of generality, we may assume $T(0) = 0$.

If $T - \nabla f_T$ is skew, then

$$\langle \nabla f_T(x) - \nabla f_T(y), x - y\rangle = \langle T(x) - T(y), x - y\rangle \geq 0,$$

so ∇f_T is monotone. By Theorem 12.17 in [12] f_T is convex, so T is decomposable. On the other hand, suppose $T = \nabla g + S$ for some convex g and skew S. Then

$$\begin{aligned} f_T(x) &= \int_0^1 \langle \nabla g(tx) + S(tx), x\rangle dt \\ &= \int_0^1 \langle \nabla g(tx), x\rangle dt = g(x) - g(0), \end{aligned}$$

so $T - \nabla f_T = T - \nabla g = S$ is skew. □

So far we have not explicitly established that (weakly) indecomposable monotone operators actually exist. We address this in the next section.

3. Indecomposable examples. The next example specifies an entire class of everywhere-defined indecomposable operators. We require the following lemma.

LEMMA 4. *Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be $\mathcal{C}^1$ and monotone. If there exist $x, y \in \mathbb{R}^n$ and $1 \le i < j \le n$ such that $DT(x)_{ij} - DT(x)_{ji} \ne DT(y)_{ij} - DT(y)_{ji}$, then T is indecomposable on $\mathbb{R}^n$.*

Proof. Suppose that $T = \nabla f + S$ with f convex and S skew. Then the Hessian matrix $\nabla^2 f(z) = DT(z) - S$ is symmetric for each $z \in \mathbb{R}^n$. Setting $\Delta_{ij} = S_{ij} - S_{ji}$, we have

$$DT(x)_{ij} = DT(x)_{ji} + \Delta_{ij} \quad \text{and} \quad DT(y)_{ij} = DT(y)_{ji} + \Delta_{ij},$$

which implies $DT(x)_{ij} - DT(x)_{ji} = DT(y)_{ij} - DT(y)_{ji}$, a contradiction. □

PROPOSITION 5. *Let $g \ge 0$ be a nonconstant and continuous real function such that either $g(x) \ge 1 = g(0)$ or $g(x) \le 1 = g(0)$. Let*

$$G(x) := \int_0^x g \quad \text{and} \quad K(x) := \int_0^x \left\{ \frac{(1+g)}{2} \right\}^2.$$

Then

(1) *$T(x, y) := (K(x) - G(y), K(y) - G(x))$ is both continuously differentiable and maximal monotone $\mathbb{R}^2$;*
(2) *T is indecomposable on $\mathbb{R}^2$.*

Proof. To check that T is monotone, we check that the symmetric part of the Jacobian DT of T is positive semidefinite as required by Fact 3. First we compute

$$DT = \begin{pmatrix} \left(\frac{1+g(y)}{2}\right)^2 & -g(y) \\ -g(x) & \left(\frac{1+g(x)}{2}\right)^2 \end{pmatrix},$$

so

$$DT_{sym} = \frac{DT + DT^T}{2} = \begin{pmatrix} \left(\frac{1+g(x)}{2}\right)^2 & -\frac{g(x)+g(y)}{2} \\ -\frac{g(x)+g(y)}{2} & \left(\frac{1+g(y)}{2}\right)^2 \end{pmatrix}.$$

Since $\left(\frac{1+g(x)}{2}\right)^2 \ge 0$, we need only check that $\operatorname{Det} DT_{sym} \ge 0$:

$$\begin{aligned} 16 \operatorname{Det} DT_{sym} &= (1+g(x))^2 (1+g(y))^2 - 4\,(g(x)+g(y))^2 \\ &= (g(x)-1)(g(y)-1)\big((g(x)+1)(g(y)+1) + 2(g(x)+g(y))\big) \\ &\ge 0. \end{aligned}$$

The maximality of T is a consequence of Example 12.7 in [12]. Lemma 4 with $i = 1, j = 2$ shows that T is indecomposable, since g is nonconstant. □

Example 6. If $g := x^2+1$ and T is constructed as in Proposition 5, then $T(x,y) = (x+1/20\,x^5+1/3\,x^3-1/3\,y^3-y, y+1/20\,y^5+1/3\,y^3-1/3\,x^3-x)$ is indecomposable. We have

$$f_T(x,y) = \frac{1}{120}\,x^6 + \frac{1}{120}\,y^6 + \frac{1}{12}\,x^4 + \frac{1}{12}\,y^4 - \frac{1}{12}\,xy^3 - \frac{1}{12}\,yx^3 + \frac{1}{2}\,x^2 - xy + \frac{1}{2}\,y^2,$$

and the Hessian of f_T is

$$\nabla^2 f_T(x,y) = \begin{bmatrix} 1/4\,x^4 + x^2 - 1/2\,xy + 1 & -1/4\,x^2 - 1/4\,y^2 - 1 \\ -1/4\,x^2 - 1/4\,y^2 - 1 & 1/4\,y^4 + y^2 - 1/2\,xy + 1 \end{bmatrix};$$

since $\nabla^2 f_T(x,y)_{11} < 0$ for large y and small positive x, f_T is not convex.

By Theorem 2, T is also not weakly decomposable.

Example 7. Consider the mapping

$$T(x,y) := \left(\sinh(x) - \alpha\,y^2/2, \sinh(y) - \alpha\,x^2/2)\right).$$

Then

$$DT = \begin{pmatrix} \cosh(x) & -\alpha y \\ -\alpha x & \cosh(y) \end{pmatrix}$$

which is monotone if and only if

$$\alpha^2 \le \frac{\cosh(x)}{x}\,\frac{\cosh(y)}{y}$$

for all $x, y > 0$. The right-hand side is a separable convex function, and is minimized at $x = y = x_0 = \coth(x_0) = 1.199678\ldots$. So T is monotone if and only if $\alpha^2 \le \sinh^2(x_0) = 2.276717\ldots$.

As before, since the difference between the off-diagonal entries of DT is nonconstant, T is indecomposable by Lemma 4.

We may now turn to the more general notion of an acyclic decomposition.

4. Acyclic decompositions. In this section, we reconstruct a modern version of a decomposition result found in [1]. We first need to recall some additional monotonicity notions. A mapping $T : \operatorname{dom} T \subset \mathbb{R}^n \to \mathbb{R}^n$ is said to be *N-monotone* for $N \ge 2$ if for every $x_1, x_2, \ldots, x_N \in \operatorname{dom} T$ we have

$$\sum_{i=1}^{N} \langle T(x_i), x_i - x_{i-1}\rangle \ge 0, \tag{4.1}$$

where $x_0 := x_N$. Note that 2-monotonicity is just monotonicity. We write $S \le_N T$ to indicate that $T = S + R$ for some N-monotone R. In particular, this means that $\operatorname{dom} T \subset \operatorname{dom} S$.

By duplicating entries in (4.1), it is easy to see that an N-monotone mapping is also M-monotone for $M \le N$; in particular, an N-monotone mapping is monotone. Asplund [1] showed that these classes are distinct via the following example.

Example 8. For $N \ge 2$ define a 2×2 matrix T_N by

$$T_N = \begin{pmatrix} \cos(\pi/N) & -\sin(\pi/N) \\ \sin(\pi/N) & \cos(\pi/N) \end{pmatrix}.$$

Then $x \to T_N(x)$ is N-monotone, but not $(N+1)$-monotone. A more explicit proof to this surprisingly difficult proposition can be found in [2, 3].

An operator that is N-monotone for every $N \geq 2$ is called *cyclically monotone* or *ω_0-monotone*. It is easy to see that subdifferential mappings are cyclically monotone; in fact, a classical result by Rockafellar [10] shows that subdifferential mappings are the only cyclically monotone mappings.

THEOREM 9 (maximal cyclic monotonicity [4, 10]). *Suppose $C : \operatorname{dom} C \subset \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is cyclically monotone. Then C has a maximal cyclically monotone extension $\widehat{C}$ of the form $\widehat{C} = \partial f$ for some proper lower semicontinuous convex function f. Furthermore,* $\operatorname{ran} \widehat{C} \subset \overline{\operatorname{conv}} \operatorname{ran} C$.

The fact that $\widehat{C}$ preserves the range of C is implicit in the proof of Theorem 1 in [10], where the convex function f is of the form $f(x) = \sup\{\langle x_\alpha^*, x\rangle - r_\alpha \mid x_\alpha^* \in \operatorname{ran} C\}$. For clarity, we prove the following lemma.

LEMMA 10. *Let X be a Banach space and let $x_\alpha^* \in X^*$ for $\alpha \in A$ and with each r_α real. Let $f(x) = \sup\{\langle x_\alpha^*, x\rangle - r_\alpha \mid \alpha \in A\}$. Then* $\operatorname{ran} \partial f \subset \overline{\operatorname{conv}}^*\{x_\alpha^* \mid \alpha \in A\}$.

Proof. Consider the convex function g defined on X^* by

$$g(x^*) := \inf\left\{\sum \lambda_{\alpha_i} r_{\alpha_i} \colon \sum \lambda_{\alpha_i} x_{\alpha_i} = x^*, \sum \lambda_{\alpha_i} = 1, \lambda_{\alpha_i} > 0\right\},$$

as we range over all finite subsets of $\{(x_\alpha^*, r_\alpha) \mid \alpha \in A\}$. It is easy to check that $g^*|_X = f$, and $f^* = g^{**}$ viewed in $\sigma(X^*, X)$. Now when $x^* \in \partial f(x)$ we have $f(x) + f^*(x^*) = \langle x^*, x\rangle$. Since $x \in \operatorname{dom} f$, we see that $g^{**}(x^*)$ is finite and we are done since $\operatorname{dom} g^{**} \subset \overline{\operatorname{dom} g}^*$.

Alternative proof. If the conclusion fails we may find $x^* \in \partial f(x)$, $\varepsilon > 0$, and $h \in X$ such that

$$\langle x^*, h\rangle > \varepsilon + \sup_{\alpha \in A} \langle x_\alpha^*, h\rangle, \tag{4.2}$$

by the Hahn–Banach theorem. Thus, for each $\alpha \in A$ we have

$$\begin{aligned}\langle x^*, h\rangle &\geq \varepsilon + \langle x_\alpha^*, x\rangle = \varepsilon + (\langle x_\alpha^*, x+h\rangle - r_\alpha) - (\langle x_\alpha^*, x\rangle - r_\alpha)\\ &\geq \varepsilon + \langle x_\alpha^*, x+h\rangle - r_\alpha - f(x).\end{aligned}$$

Now $f(x)$ is finite and so supremizing over $\alpha \in A$ yields

$$\langle x^*, h\rangle \geq \varepsilon + f(x+h) - f(x),$$

in contradiction to $x^* \in \partial f(x)$. ∎

Another range-preserving extension theorem we shall require is the following central case of the Debrunner–Flor theorem.

THEOREM 11 (Debrunner–Flor extension [4, 6]). *Suppose $T : \operatorname{dom} T \subset \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is monotone with range in MB_{X^*} for some $M > 0$. Then T has a bounded monotone extension $\widehat{T}$ with* $\operatorname{dom} \widehat{T} = \mathbb{R}^n$ *and* $\operatorname{ran} \widehat{T} \subset \overline{\operatorname{conv}} \operatorname{ran} T$.

The proof of the decomposition below hinges on a kind of monotone convergence theorem. We require the following definition: a monotone operator T is *3^--monotone* if

$$\langle T(x), y\rangle \leq \langle T(x), x\rangle + \langle T(y), y\rangle$$

for all $x, y \in \operatorname{dom} T$. In particular, this holds if T is N-monotone for $N \geq 3$, and $0 \in T(0)$.

THEOREM 12 (monotone convergence [1, 4]). *Let N be one of $3^-, 3, 4, \dots,$ or ω_0. Consider an increasing net of monotone operators $T_\alpha : \operatorname{dom} T_\alpha \subset \mathbb{R}^n \to \mathbb{R}^n$ satisfying*

$$0 \leq_N T_\alpha \leq_N T_\beta \leq_2 T,$$

whenever $\alpha < \beta \in \mathcal{A}$, for some monotone $T : \operatorname{dom} T \subset \mathbb{R}^n \to \mathbb{R}^n$. Suppose that $T(0) = 0$, $T_\alpha(0) = 0$ for all α, and that $0 \in \operatorname{int} \operatorname{dom} T$. Then

(i) *there is an N-monotone operator $T_{\mathcal{A}}$ with*

$$T_\alpha \leq_N T_{\mathcal{A}} \leq_2 T$$

for all $\alpha \in \mathcal{A}$;

(ii) *if T is maximal monotone and $\operatorname{ran} T \subset MB$ for some $M > 0$, then one may assume $\operatorname{ran} T_{\mathcal{A}} \subset MB$.*

Proof. (i) Let $\alpha < \beta$. Since $T(0) = 0$ and $0 \leq_2 T_\alpha \leq_2 T_\beta \leq_2 T$, we have

$$0 \leq \langle x, T_\alpha(x)\rangle \leq \langle x, T_\beta(x)\rangle \leq \langle x, T(x)\rangle \tag{4.3}$$

for $x \in \operatorname{dom} T$. So $\lim_{\alpha\to\infty}\langle x, T_\alpha(x)\rangle$ exists.

Writing $T_{\beta\alpha} = T_\beta - T_\alpha$ and using $T_{\beta\alpha} \geq_{3^-} 0$, we get

$$\langle y, T_{\beta\alpha}(x)\rangle \leq \langle x, T_{\beta\alpha}(x)\rangle + \langle y, T_{\beta\alpha}(y)\rangle \tag{4.4}$$

for $x, y \in \operatorname{dom} T$. A monotone operator is locally bounded on the interior of its domain (see [4]) and $0 \in \operatorname{int} \operatorname{dom} T$, so there exist $\varepsilon > 0$ and $M > 0$ with $T(\varepsilon B) \subset MB$ and $\varepsilon B \subset \operatorname{dom} T$. Then

$$0 \leq \langle y, T_{\beta\alpha}(y)\rangle \leq \langle y, T(y)\rangle \leq \varepsilon M \tag{4.5}$$

when $\|y\| \leq \varepsilon$.

For $x \in \operatorname{dom} T$, we may choose $\gamma(x)$ so that

$$0 \leq \langle x, T_{\beta\alpha}(x)\rangle \leq \varepsilon^2 \tag{4.6}$$

whenever $\beta > \alpha > \gamma(x)$, since $\langle x, T_\alpha(x)\rangle$ is convergent.

Combining (4.4), (4.5), and (4.6) gives

$$\langle y, T_{\beta\alpha}(x)\rangle \leq \langle x, T_{\beta\alpha}(x)\rangle + \langle y, T_{\beta\alpha}(y)\rangle \leq (M + \varepsilon)\varepsilon$$

for all $\|y\| \leq \varepsilon$ and $\beta > \alpha > \gamma(x)$. This shows that

$$\langle y, T_{\beta\alpha}(x)\rangle \to 0$$

for all $y \in \mathbb{R}^n$, so $\big(T_\alpha(x)\big)_\alpha$ is Cauchy, and thus has a limit. Setting $T_{\mathcal{A}}(x)$ to this limit, it is clear from the definitions that $T_{\mathcal{A}}$ is N-monotone. It is straightforward to check $T_\alpha \leq_N T_{\mathcal{A}} \leq_2 T$.

(ii) The Debrunner–Flor result shows that $\operatorname{dom} T = \mathbb{R}^n$, since T is maximal. Fixing $x \in \mathbb{R}^n$, we know that

$$\begin{aligned}\langle T_\alpha(x), y\rangle &\leq \langle T_\alpha(x), x\rangle + \langle T_\alpha(y), y\rangle \\ &\leq \langle T(x), x\rangle + \langle T(y), y\rangle\end{aligned}$$

for all $y \in \operatorname{dom} T = \mathbb{R}^n$.

From $\|T(y)\| \le M$ we get

$$\|T_\alpha(x)\|\ \|y\| \le \langle T(x), x\rangle + M\|y\|$$

for all $y \in \mathbb{R}^n$. Letting $\|y\| \to \infty$ in this expression gives $\|T_\alpha(x)\| \le M$. □

The maximality condition in part (ii) of Theorem 12 cannot be removed for $N \neq \omega_0$. Indeed, for a fixed $N \ge 3$ and T_N as in Example 8, define maps T_α and T on the unit ball B by $T_\alpha(x) := T_N(x)$ for each α in some directed set $\mathcal{A}$ and $T(x) := \left(\frac{T_N + T_N^T}{2}\right)x = \cos(\pi/N)Ix$. Then $0 \le_N T_\alpha \le T_\beta \le T$ for $\alpha < \beta$, and $T_{\mathcal{A}} = T_\alpha$, but

$$\operatorname{ran} T_{\mathcal{A}} = T_{\mathcal{A}}(B) = B \nsubseteq \cos(\pi/N)B = \operatorname{ran} T.$$

Now we are ready to present an updated version of a decomposition result provided in [1]. In this case, the decomposition takes the form of a subdifferential component, as before, and an *acyclic* (termed *irreducible* in [1]) remainder A.

Given a set $C \subset \mathbb{R}^n$, a monotone operator $A : \mathbb{R}^n \to \mathbb{R}^n$ is said to be *acyclic* with respect to C if $A = \partial f + R$ with R monotone implies that ∂f is constant on C (i.e., f is affine on C). That is, $A|_C$ has no nontrivial subdifferential component. If no set C is given, then $C = \operatorname{dom} A$ is implied.

THEOREM 13 (Asplund decomposition [1, 4]). *Suppose we are given a (single-valued) maximal monotone operator* $T : \operatorname{dom} T \subset \mathbb{R}^n \to \mathbb{R}^n$ *with* $\operatorname{int}\operatorname{dom} T \neq \emptyset$.

(i) *T may be decomposed as*

$$T = \nabla f + A,$$

where f is lower semicontinuous and convex, while A is acyclic with respect to $\operatorname{dom} T$.

(ii) *If* $\operatorname{ran} T \subset MB$, *we may assume that f is M-Lipschitz.*

Proof. (i) First, shift the graph of T so that $0 \in \operatorname{int}\operatorname{dom} T$. Consider the set

$$\mathcal{C} := \{C \mid 0 \le_{\omega_0} C \le_2 T,\ C(0) = 0\},$$

ordered by the partial order $\le_{\omega_0}$. Every chain in $\mathcal{C}$ has an upper bound $T_{\mathcal{A}}$ by Theorem 12, and $\mathcal{C}$ is nonempty since it contains the zero mapping, so Zorn's lemma provides a $\le_{\omega_0}$-maximal $\widehat{C}$ in $\mathcal{C}$ with

$$0 \le_{\omega_0} \widehat{C} \le_2 T.$$

So $T = \widehat{C} + A$ for some monotone A. To show that A is acyclic, suppose $A = \partial g + M$. Then

$$T = (\widehat{C} + \partial g) + M,$$

so, by adding a constant to ∂g and subtracting it from M if necessary, we have $\partial g + \widehat{C} \in \mathcal{C}$. Since $\widehat{C}$ is $\le_{\omega_0}$-maximal, we have $\widehat{C} + \partial g \le_{\omega_0} \widehat{C}$, so $\operatorname{gr}(-\partial g|_{\operatorname{dom} T}) \subset \operatorname{gr}\partial h$ for some lower semicontinuous convex $h : \mathbb{R}^n \to \mathbb{R}$. Thus g is both convex and concave, hence affine, on $\operatorname{dom} T$, and A is therefore acyclic with respect to $\operatorname{dom} T$.

Now, $\widehat{C}$ is cyclically monotone, so Rockafellar's result shows that $\operatorname{gr}\widehat{C} \subset \operatorname{gr}\partial f$ for some proper convex lower semicontinuous f. This gives

$$\operatorname{gr} T = \operatorname{gr}(\widehat{C} + A) \subset \operatorname{gr}(\partial f + A),$$

but $\partial f + A$ is monotone, and T is maximal monotone, so $T = \partial f + A$, as required.

(ii) Part (ii) of Theorem 12 shows that one may assume that $\operatorname{ran} \widehat{C} \subset MB$, so $\operatorname{ran} \partial f \subset MB$ by Rockafellar's result. It is straightforward to show that this implies that f is M-Lipschitz. ∎

An immediate corollary of this decomposition is the following.

COROLLARY 14 (nonlinear acyclicity). *Under the hypotheses of Theorem* 13*, if* $T : \operatorname{dom} T \subset \mathbb{R}^n \to \mathbb{R}^n$ *is maximal monotone with bounded range, then the acyclic part of the Asplund decomposition of* T *is either nonlinear or zero. In other words,* A *is nonlinear unless* T *is cyclically monotone.*

Proof. Since T is maximal monotone with bounded range, $\operatorname{dom} T = \mathbb{R}^n$. The decomposition $T = \partial f + A$ shows that $\operatorname{dom} \partial f = \operatorname{dom} A = \mathbb{R}^n$, and we know that the range of ∂f is bounded as well. If A is nonzero and linear, then the range of A is unbounded, which is impossible. ∎

Corollary 14 immediately implies the existence of many nonlinear acyclic operators, but it does not exhibit any explicitly. We remedy this in the next and final section.

5. Explicit acyclic examples. Skew linear mappings are canonical examples of monotone mappings that are not subdifferential mappings. It is therefore reassuring to know that they are acyclic.

PROPOSITION 15. *Suppose that* $S : \mathbb{R}^n \to \mathbb{R}^n$ *is a continuous linear operator satisfying* $\langle S(x), x\rangle = 0$ *for all* $x \in \mathbb{R}^n$*. Then* S *is acyclic.*

Proof. Let $S = F + R$, where F is a subdifferential mapping and R is maximal monotone. Since S is single valued, F and R are single valued. In particular, $F = \nabla f$ for some convex differentiable f. Since R is monotone, we have

$$\begin{aligned} 0 &\le \langle R(x) - R(y), x - y\rangle = \langle S(x) - S(y), x - y\rangle - \langle F(x) - F(y), x - y\rangle \\ &= -\langle F(x) - F(y), x - y\rangle = \langle \nabla(-f)(x) - \nabla(-f)(y), x - y\rangle. \end{aligned}$$

This shows that $-f$ is convex, so f is convex and concave, hence linear on its domain. But $\operatorname{dom} f \supset \operatorname{dom} S = \mathbb{R}^n$, so $f \in \mathbb{R}^n$. So $F = \nabla f$ is constant. In fact, by subtracting from F and adding to R, we may assume that $F = 0$. ∎

We leave it to the reader to check that the sum of an acyclic operator and a skew linear operator is still acyclic. It is not clear that the sum of two acyclic operators must be acyclic. For continuous linear monotone operators, then, the usual decomposition into symmetric and skew parts is the same as the Asplund decomposition into subdifferential and acyclic parts.

We recall that Asplund was unable to find explicit examples of nonlinear acyclic mappings [1], and we have found this quite challenging as well. In particular, we wish to determine a useful characterization of acyclicity. We make some progress in this direction by providing an explicit and, to our mind, surprisingly simple example: we present a nonlinear acyclic monotone mapping $\widehat{S} : \mathbb{R}^2 \to \mathbb{R}^2$.

Precisely, $\widehat{S}$ is constructed by restricting the range of the skew mapping $S(x, y) = (-y, x)$ to the unit ball and taking a range-preserving maximal monotone extension of the restriction. This extension is unique, as we see from the following corollary of Proposition 14 from [4], work that originates in [7].

COROLLARY 16 (unique extension [4, 7]). *Suppose* $T : \mathbb{R}^n \to \mathbb{R}^n$ *is maximal monotone and suppose that* $\operatorname{ran} T \cap \operatorname{int} B \neq \emptyset$*. Then there is a unique maximal monotone mapping* $\widehat{T}$ *such that* $T(x) \cap B \subset \widehat{T}(x) \subset B$*. Furthermore,*

$$\widehat{T}(x) = \{x^* \in B \mid \langle x^* - y^*, x - y\rangle \ge 0 \text{ for all } y^* \in T(y) \cap \operatorname{int} B\}. \tag{5.1}$$

Note that $\widehat{T}$ is either a Lipschitz subgradient or it has a nonlinear acyclic part: the acyclic part is bounded so it cannot be nontrivially linear. Hence in the construction of Proposition 17 we know that $\widehat{S}$ has nonlinear acyclic part, which we shall eventually show in Proposition 20 to be $\widehat{S}$ itself.

PROPOSITION 17. *Define $S : \mathbb{R}^2 \to \mathbb{R}^2$ by $S(x, y) = (-y, x)$ for $x^2 + y^2 \leq 1$. Then the unique maximal monotone extension $\widehat{S}$ of S with range restricted to the unit disc is*

$$\widehat{S}(x) = \begin{cases} S(x), & \|x\| \leq 1, \\ \sqrt{1 - \frac{1}{\|x\|^2}}\,\frac{x}{\|x\|} + \frac{1}{\|x\|} S\left(\frac{x}{\|x\|}\right), & \|x\| > 1. \end{cases}$$

Proof. From Corollary 16, we know that $\widehat{S}$ exists and is uniquely defined. In the interior of the unit ball, (5.1) shows that $\widehat{S}(x) = S(x)$. Indeed, let $t > 0$ be so small that $z = x + ty \in B$ for all unit length y. Then

$$\langle S(x + ty) - \widehat{S}(x), y\rangle \geq 0$$

for all unit y. Letting $t \to 0$ shows that $\widehat{S}(x) = S(x)$. To determine $(u, v) = \widehat{S}(x)$ for $\|x\| \geq 1$, it suffices by rotational symmetry to consider points $x = (a, 0)$ with $a \geq 1$. Then monotonicity requires that

$$\langle \widehat{S}(x) - S(z), x - z\rangle \geq 0$$

for all $\|z\| \leq 1$. Let $z = \left(\frac{1}{a}, -\frac{\sqrt{a^2-1}}{a}\right)$ so that $\widehat{S}(z) = S(z) = \left(\frac{\sqrt{a^2-1}}{a}, \frac{1}{a}\right)$. Then

$$\left\langle (u, v) - \left(\frac{\sqrt{a^2 - 1}}{a}, \frac{1}{a}\right), (a, 0) - \left(\frac{1}{a}, -\frac{\sqrt{a^2 - 1}}{a}\right)\right\rangle \geq 0.$$

Expanding this gives

$$u\left(a - \frac{1}{a}\right) + \sqrt{1 - \frac{1}{a^2}}(v - a) \geq 0,$$

and noting that $u \leq \sqrt{1 - v^2}$ gives

$$\sqrt{1 - v^2}(a^2 - 1) + \sqrt{a^2 - 1}(v - a) \geq 0$$

which reduces to $(av - 1)^2 \leq 0$, that is, $v = 1/a$. Similarly, setting $z = (\frac{1}{a}, -\frac{\sqrt{a^2-1}}{a})$ also shows that $u = \sqrt{1 - 1/a^2}$.

So

$$\widehat{S}(x) = \widehat{S}(a, 0) = \left(\sqrt{1 - \frac{1}{a^2}}, \frac{1}{a}\right) = \sqrt{1 - \frac{1}{\|x\|^2}}\,\frac{x}{\|x\|} + \frac{1}{\|x\|} S\left(\frac{x}{\|x\|}\right).$$

The same result holds for general $\|x\| \geq 1$ by considering the coordinate system given by the orthogonal basis $\{x, S(x)\}$. ∎

Figure 1 shows the graph of the vector field $\widehat{S}$. Having computed $\widehat{S}$, we commence to show that it is acyclic, with the aid of two technical lemmas.

LEMMA 18. *$\widehat{S}(x + tS(x)) = S(x)$ for all $t \geq 0$ and all $\|x\| = 1$.*

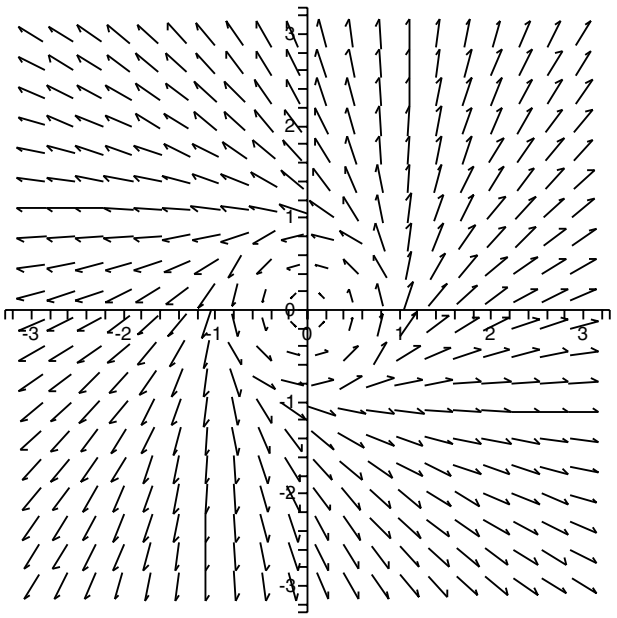


FIG. 1. *A field plot of* $\widehat{S}$.

Proof.

$$\begin{aligned}\widehat{S}(x+tS(x)) &= \sqrt{1-\frac{1}{1+t^2}}\,\frac{x+tS(x)}{\sqrt{1+t^2}} + \frac{1}{1+t^2}\,S(x+tS(x)) \\ &= \frac{t}{1+t^2}\,(x+tS(x)) + \frac{1}{1+t^2}\,(S(x)-tx) = S(x),\end{aligned}$$

since $S^2 = -I$. ∎

This construction does not extend immediately to all skew mappings, since it assumes that $S^2 = -I$, which can occur only in even dimensions.

FACT 4. *Skew orthogonal matrices exist only in even dimensions.*

Proof. $\operatorname{Det} S = \operatorname{Det}(S^\top) = \operatorname{Det}(-S) = (-1)^n \operatorname{Det} S$. ∎

However, such mappings do exist for each even-dimensional $\mathbb{R}^{2n}$, and these can be embedded in $\mathbb{R}^{2n+1}$ in an obvious way. Thus, our construction provides an acyclic nonlinear mapping for each $\mathbb{R}^n$, $n > 1$.

To show that $\widehat{S}$ is acyclic, we suppose that $\widehat{S} = F + R$, where $F = \partial f$ for some convex proper lower semicontinuous function f and R is maximal monotone, and we show that F is constant.

LEMMA 19. *Let* $\|x\| = 1$, $t \geq 0$, *and* $y(t) = x + tS(x)$. *Then* $\langle F(y(t)), S(x)\rangle = c(x)$ *for some constant* $c(x)$.

Proof. Suppose $t_1 \neq t_2$. Then $\widehat{S}(y(t_1)) = \widehat{S}(y(t_2)$, by Lemma 18, so

$$\begin{aligned}0 &\leq \langle R(y(t_1)) - R(y(t_2)),\, y(t_1) - y(t_2)\rangle \\ &= \langle \widehat{S}(y(t_1)) - \widehat{S}(y(t_2)),\, y(t_1) - y(t_2)\rangle - \langle F(y(t_1)) - F(y(t_2)),\, y(t_1) - y(t_2)\rangle \\ &= -\langle F(y(t_1)) - F(y(t_2)),\, y(t_1) - y(t_2)\rangle \leq 0,\end{aligned}$$

and so

$$\langle F(y(t_1)) - F(y(t_2)), x + t_1 S(x) - (x + t_2 S(x))\rangle = 0;$$

that is,

$$\langle F(y(t_1)), S(x)\rangle = \langle F(y(t_2)), S(x)\rangle$$

for any t_1, t_2. □

PROPOSITION 20. *The extension mapping $\widehat{S}$ given explicitly in Proposition* 17 *is nonlinear and acyclic with bounded range and full domain.*

Proof. First note that if $\widehat{S} = F + R$ with R monotone and $F = \partial f$, then both are single valued, so $F = \nabla f$. As in Proposition 15, we can assume that $f(x) = 0$ when $\|x\| \leq 1$.

Let $\|y\| > 1$. Then there are a unit vector x and a t such that $y = x + tS(x)$:

$$x = \widehat{x}(y) := \frac{y}{\|y\|^2} - \sqrt{\frac{1}{\|y\|^2} - \frac{1}{\|y\|^4}}\, S(y),$$

$$t = t(y) = \sqrt{\|y\|^2 - 1},$$

and we note that $y \to \widehat{x}(y)$ is continuous. We will determine $f(y)$ by integrating F along the ray $s \to x + sS(x)$. Using Lemma 19, we have

$$\begin{aligned}f(y) - f(x) &= \int_0^t \langle \nabla f(x + sS(x)), S(x)\rangle ds \\ &= \int_0^t c(x) ds = c(x)t.\end{aligned}$$

Since f is continuous and convex, c is continuous and positive, so $y \to c(\widehat{x}(y))$ is continuous and positive.

Plugging in $t(y)$ gives $f(y) = c(\widehat{x}(y))\sqrt{\|y\|^2 - 1}$ when $\|y\| > 1$ and $f = 0$ for $\|y\| \leq 1$. Suppose $c(y) > 0$ for some $\|y\| = 1$. Then for f to be convex on the segment $[y, 2y]$ we require that

$$(1 - \lambda) f(y) + \lambda f(2y) \geq f\big((1 + \lambda)y\big) \quad \text{for all } \lambda \in (0, 1).$$

This means that

$$0 + \lambda c\big(\widehat{x}(2y)\big)\sqrt{3} \geq c\big(\widehat{x}\big((1 + \lambda)y\big)\big)\sqrt{\lambda^2 + 2\lambda}$$

or

$$c\big(\widehat{x}(2y)\big)\sqrt{3} \geq c\big(\widehat{x}\big((1 + \lambda)y\big)\big)\sqrt{1 + \frac{2}{\lambda}}$$

for all $\lambda \in (0,1)$. Letting $\lambda \to 0$, we get $\widehat{x}\big((1+2\lambda)y\big) \to y$, so $c\big(\widehat{x}\big((1+\lambda)y\big)\big) \to c(y) > 0$. Since $\sqrt{1+2/\lambda} \to \infty$, the inequality does not hold for small λ unless $c(y) = 0$.

For f to be convex and everywhere defined, then, we require $c(y) = 0$ for all $\|y\| = 1$. That is, f is identically zero. ∎

It seems probable that the construction above applied to any nontrivial skew linear mapping always leads to an acyclic mapping—and that more ingenuity will allow some reader to prove this. We conclude the paper by exploring Fitzpatrick's last function for $\widehat{S}$ as above.

6. Computing $f_{\widehat{S}}$. We can also explicitly compute Fitzpatrick's last function $f_{\widehat{S}}$ as in the previous section. We have the following proposition.

PROPOSITION 21. *With $\widehat{S}$ as before, we have*

$$f_{\widehat{S}}(x) = \begin{cases} 0, & \|x\| \le 1, \\ \sqrt{\|x\|^2 - 1} + \arctan\left(\frac{1}{\sqrt{\|x\|^2-1}}\right) - \frac{\pi}{2}, & \|x\| > 1. \end{cases}$$

Proof. It is immediate from the definition that $f_{\widehat{S}}(x) = 0$ when $\|x\| \le 1$. For $\|x\| > 1$, we get

$$\begin{aligned} f_{\widehat{S}}(x) &= \int_0^1 \langle x, \widehat{S}(tx)\rangle dt \\ &= \int_0^{\frac{1}{\|x\|}} t\langle x, S(x)\rangle dt + \int_{\frac{1}{\|x\|}}^1 \sqrt{1 - \frac{1}{t^2\|x\|^2}}\,\frac{1}{\|x\|}\langle x, x\rangle dt + \int_{\frac{1}{\|x\|}}^1 \frac{1}{t\|x\|^2}\langle S(x), x\rangle dt \\ &= \int_{\frac{1}{\|x\|}}^1 \sqrt{1 - \frac{1}{t^2\|x\|^2}}\,\|x\| dt \\ &= \int_1^{\|x\|} \sqrt{1 - \frac{1}{s^2}}\, ds \\ &= \sqrt{\|x\|^2 - 1} + \arctan\left(\frac{1}{\sqrt{\|x\|^2 - 1}}\right) - \frac{\pi}{2}. \end{aligned}$$ ∎

Note that $f_{\widehat{S}}$ is convex, since it is a composition of the norm $x \to \|x\|$ with the increasing convex function $t \to \int_1^t \sqrt{1 - 1/s^2}\, ds$. So $\widehat{S}$ is weakly decomposable as $\widehat{S} = \nabla f_{\widehat{S}} + SL$ where SL is skew-like. To determine SL, we compute

$$\nabla f_{\widehat{S}}(x) = \begin{cases} 0, & \|x\| < 1, \\ \sqrt{1 - \frac{1}{\|x\|^2}}\,\frac{x}{\|x\|}, & \|x\| \ge 1. \end{cases}$$

So $\widehat{S}(x) = \nabla f_{\widehat{S}}(x) + h(\|x\|)S(x)$, where

$$h(t) = \begin{cases} 1, & t \le 1, \\ \frac{1}{t^2}, & t \ge 1. \end{cases}$$

So $\widehat{S}$ is not decomposable, but is weakly decomposable, since $SL = x \to h(\|x\|)S(x)$ is clearly skew-like. Note finally that SL is not monotone.

7. Conclusion. In this paper, we have provided some tools for the decomposition of monotone operators. This was originally motivated by observing that the classical counterexamples in monotone operator theory (see section 6 of [4]) are built from skew operators; in some sense, subgradients ("symmetric" operators) and acyclic mappings ("skew" operators) represent the extreme points of the space of monotone operators. The results we have given in this paper make this more concrete.

We remain convinced that a better understanding of acyclic operators will shed light on a number of open questions. For instance, if a Banach space has good differentiability properties, do all monotone operators defined on the space inherit these properties? Are such properties determined by the behavior of the acyclic part? In a more limited fashion it seems important to answer the following questions: (1) Is there an iterative construction to compute the acyclic part of a monotone operator in finite-dimensional space? (2) Is there an effective characterization of acyclicity that allows one to easily determine whether a given operator is acyclic? (3) When is the sum of acyclic mappings acyclic? (4) Can one exhibit an acyclic mapping whose domain is not the whole space?

Acknowledgments. We would like to thank Heinz Bauschke and the referees whose very careful reading greatly improved the paper's exposition.

REFERENCES

[1] E. Asplund, *A monotone convergence theorem for sequences of nonlinear mappings*, Proc. Sympos. Pure Math., 18 (1970), pp. 1–9.
[2] S. Bartz, H. H. Bauschke, J. M. Borwein, S. Reich, and X. Wang, *Fitzpatrick functions, cyclic monotonicity, and Rockafellar's antiderivative*, Nonlinear Anal., 66 (2007), pp. 1198–1223.
[3] H. H. Bauschke, J. M. Borwein, and X. Wang, *Fitzpatrick functions and continuous linear monotone operators*, SIAM J. Optim., 18 (2007), pp. 789–809.
[4] J. M. Borwein, *Maximal monotonicity via convex analysis*, J. Convex Anal., 13 (2006), pp. 561–586.
[5] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Canadian Mathematical Society Series of Monographs and Advanced Texts, John Wiley and Sons, New York, 1983.
[6] H. Debrunner and P. Flor, *Ein Erweiterungssatz für monotone Mengen*, Arch. Math., 15 (1964), pp. 445–447.
[7] S. Fitzpatrick and R. R. Phelps, *Bounded approximants to monotone operators on Banach spaces*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 9 (1992), pp. 573–595.
[8] S. Fitzpatrick and R. R. Phelps, *Some properties of maximal monotone operators on nonreflexive Banach spaces*, Set-Valued Anal., 3 (1995), pp. 51–69.
[9] R. R. Phelps, *Convex Functions, Monotone Operators, and Differentiability*, Lecture Notes in Math. 1364, Springer-Verlag, Berlin, 1989.
[10] R. T. Rockafellar, *Characterization of the subdifferentials of convex functions*, Pacific J. Math., 17 (1966), pp. 497–510.
[11] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[12] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Grundlehren Math. Wiss. 317, Springer-Verlag, Berlin, Heidelberg, New York, 1998.
[13] S. Simons, *Subdifferentials are locally maximal monotone*, Bull. Austral. Math. Soc., 47 (1993), pp. 465–471.
[14] S. Simons, *Minimax and Monotonicity*, Lecture Notes in Math. 1693, Springer-Verlag, New York, 1998.

SIAM J. OPTIM.
Vol. 18, No. 3, pp. 961–979

STABILITY OF ε-APPROXIMATE SOLUTIONS TO CONVEX STOCHASTIC PROGRAMS*

W. RÖMISCH† AND R. J.-B. WETS‡

Abstract. An analysis of convex stochastic programs is provided when the underlying probability distribution is subjected to (small) perturbations. It is shown, in particular, that ε-approximate solution sets of convex stochastic programs behave Lipschitz continuously with respect to certain distances of probability distributions that are generated by the relevant integrands. It is shown that these results apply to linear two-stage stochastic programs with random recourse. We discuss the consequences on associating Fortet–Mourier metrics to two-stage models and on the asymptotic behavior of empirical estimates of such models, respectively.

Key words. stochastic programming, quantitative stability, approximate solutions, probability metrics, two-stage models, random recourse

AMS subject classifications. 90C15, 90C31

DOI. 10.1137/060657716

1. Introduction.

Stochastic programming deals with models for optimization problems under (stochastic) uncertainty that require a decision on the basis of probabilistic information about random data. Typically, deterministic equivalents of such models are finite- or infinite-dimensional nonlinear programs depending on the properties of the distribution of the random components of the problems. Their solutions depend on the probability distribution of the random data via certain expectation functionals. Many deterministic equivalents of stochastic programming models take the form

$$\min\left\{\mathbb{E}^P f_0(x) := \int_{\Xi} f_0(\xi, x)\, P(d\xi) : x \in X\right\}, \tag{1.1}$$

where X is a closed convex subset of $\mathbb{R}^m$, Ξ is a closed subset of $\mathbb{R}^s$, P is a Borel probability measure on Ξ, and $\mathbb{E}^P$ denotes expectation with respect to P. The function f_0 from $\mathbb{R}^m \times \Xi$ to $\overline{\mathbb{R}} = [-\infty, \infty]$ is a *convex random lower semicontinuous (lsc) function*,[1] and, in particular, this means

- $(\xi, x) \mapsto f_0(\xi, x)$ is Borel measurable, and
- for all $\xi \in \Xi$, $f_0(\xi, \cdot)$ is lsc and convex.

It is part of the stochastic programming folklore, repeatedly observed in practice, that the solutions, or at least the approximating solutions, are quite robust with respect to reasonable perturbations of the probability distribution of the random components of the problem. In this paper, we substantiate this belief by focusing

*Received by the editors April 20, 2006; accepted for publication (in revised form) April 5, 2007; published electronically October 4, 2007. This work was supported by the DFG Research Center MATHEON *Mathematics for key technologies* in Berlin and by an NSF grant of the second author.
http://www.siam.org/journals/siopt/18-3/65771.html

†Humboldt-University Berlin, Institute of Mathematics, D-10099 Berlin, Germany (romisch@math.hu-berlin.de).

‡University of California at Davis, Department of Mathematics, Davis, CA 95616-8633 (rjbwets@ucdavis.edu).

[1] The concept of a random lsc function is due to Rockafellar [21], who introduced it in the context of the calculus of variations under the name of "normal integrand." Further properties of random lsc functions are set forth in [23, Chapter 14], [33].

our analysis on the approximating solutions for which we are able to derive Lipschitz continuity without even requiring fixed (deterministic) recourse.

In the following, we denote by $\mathcal{P}(\Xi)$ the set of all Borel probability measures on Ξ and by $v(P)$, $S(P)$, and $S_\varepsilon(P)$ ($\varepsilon \geq 0$) the infimum, the solution set, and the set of ε-approximate solutions to (1.1), i.e.,

$$\begin{aligned} v(P) &:= \inf \mathbb{E}^P f_0 := \inf \left\{ \mathbb{E}^P f_0(x) : x \in X \right\}, \\ S_\varepsilon(P) &:= \varepsilon\text{-argmin}\, \mathbb{E}^P f_0 := \left\{ x \in X \ : \mathbb{E}^P f_0(x) \leq v(P) + \varepsilon \right\}, \\ S(P) &:= \operatorname{argmin}\, \mathbb{E}^P f_0 := S_0(P). \end{aligned}$$

Since, in practice, the underlying probability distribution P is often not known precisely, the stability behavior of the stochastic program (1.1) when changing (perturbing, estimating, approximating) P is important. Here, stability refers to continuity properties of the optimal value function $v(.)$ and of the set-valued mapping $S_\varepsilon(.)$ at P, where both $v(.)$ and $S_\varepsilon(.)$ are regarded as mappings given on certain subsets of $\mathcal{P}(\Xi)$ equipped with some probability (semi)metric.

Early work on stability of stochastic programs is reported in [11, 19, 27] and later in [1]. Quantitative stability of two-stage models was studied, e.g., in [25, 26, 29, 18]. A recent survey of stability results in stochastic programming is given in [24]. Most of the recent contributions to (quantitative) stability use the general framework and the results of [3, 14] and [23, Chapter 7J], respectively.

In the present paper, we take up an issue brought to the fore in [38, section 4]. Since solutions derived, when actually solving (1.1), are usually ε-approximate solutions of an approximating problem where P has been replaced by an approximating measure Q, it is crucial to investigate the (quantitative) continuity properties of the (set-valued) mapping ε-argmin as a function of P, i.e., $P \mapsto S_\varepsilon(P)$, from $\mathcal{P}$ of probability measures to the space of closed convex subsets of $\mathbb{R}^m$.

Quantitative perturbation results for ε-approximate solutions in optimization are given in [4] and [23, Chapter 7J]. The corresponding estimates make use of the epi-distance between the objective functions of (1.1) and its perturbations. In our analysis, the corresponding subset $\mathcal{P}$ of probability measures is determined by satisfying certain moment conditions that are related to growth properties of the integrand f_0 with respect to ξ. The epi-distances of the objective functions can be bounded by some probability semimetric of the form

$$d_{\mathcal{F}}(P,Q) = \sup \left\{ \left| \int_\Xi f(\xi) P(d\xi) - \int_\Xi f(\xi) Q(d\xi) \right| \ : \ f \in \mathcal{F} \right\}, \tag{1.2}$$

where $\mathcal{F}$ is an appropriate class of measurable functions from Ξ to $\overline{\mathbb{R}}$ and P, Q are probability measures in $\mathcal{P}$. First, we show in section 2 that classes of the form $\mathcal{F}_\rho = \{f_0(\cdot, x) : x \in X \cap \rho\mathbb{B}\}$ for some $\rho > 0$ and $\mathbb{B}$ denoting the unit ball in $\mathbb{R}^m$ and the corresponding distance $d_{\mathcal{F}_\rho}$ are suitable to derive the desired stability results.

In section 3 we then provide characterizations of the function classes $\mathcal{F}_\rho$ for two-stage models with random recourse. Two-stage stochastic programs arise as deterministic equivalents of improperly posed random linear programs of the form

$$\min\{cx : x \in X, T(\xi)x = h(\xi)\},$$

where X is polyhedral and the (technology) matrix $T(\xi)$ and the vector $h(\xi)$ depend on a random vector ξ. Given a realization of ξ, a possible deviation $h(\xi) - T(\xi)x$

is compensated for by the additional cost $q(\xi)y(\xi)$, where $y = y(\xi)$ belongs to a polyhedral set Y and satisfies $W(\xi)y = h(\xi) - T(\xi)x$. Here, the cost coefficient $q(\xi)$ and the compensation or recourse matrix $W(\xi)$ (may) depend on the realization. The modeling idea consists in adding the expected compensation cost $\mathbb{E}[q(\xi)y(\xi)]$ to cx. By minimizing the objective function $cx + \mathbb{E}[q(\xi)y(\xi)]$ first with respect to $y(\xi)$, we arrive at the function

$$f_0(\xi, x) := cx + \inf\{q(\xi)y : y \in Y, W(\xi)y = h(\xi) - T(\xi)x\},$$

whose expectation has to be minimized with respect to $x \in X$. Since the decisions x and $y(\xi)$ are made before or after the realization of ξ, they are called first- and second-stage decisions, respectively.

While Lipschitz continuity properties of the integrands f_0 with respect to ξ are well understood for fixed recourse [36], much less is known for random recourse. In section 3 we deal with the following two cases: (i) full random recourse by imposing local Lipschitz continuity of the (second-stage) dual feasibility mapping and (ii) a specific lower diagonal randomness of the recourse matrix. The latter situation occurs, for example, in the following two important cases.

Let us first consider a dynamical decision process, as in a variety of applications, where the compensation idea is repeated l times after the realization of a new random vector ξ_j, $j = 1, \dots, l$. Then we have second-stage decisions $y_j = y_j(\xi_j)$ with corresponding cost $q_j(\xi_j)y_j$ which satisfy the constraints $y_j \in Y_j$ and $W_{jj}y_j = h_j(\xi_j) - W_{jj-1}(\xi_j)y_{j-1}$ for $j = 1, \dots, l$, where $l \in \mathbb{N}$, y_0 is the first-stage decision and $W_{jj-1}(\xi_j)$ are (random) technology matrices. This leads to the function

$$f_0(\xi, y_0) = cy_0 + \inf\left\{\sum_{j=1}^{l} q_j(\xi)y_j : W_{jj}y_j = h_j(\xi) - W_{jj-1}(\xi)y_{j-1}, y_j \in Y_j, j = 1, \dots, l\right\},$$

where $\xi = (\xi_1, \dots, \xi_l)$ and $q_j(\xi) := q_j(\xi_j)$, etc. The expectation of this function is to be minimized in multiperiod two-stage stochastic programming models. If we introduce the second-stage decision vector $y = (y_1, \dots, y_l)$, the corresponding recourse matrix $W(\xi)$ is a block lower triangular matrix containing W_{jj}, $j = 1, \dots, l$, in the main diagonal and $W_{jj-1}(\xi)$, $j = 1, \dots, l$, in the lower diagonal (see section 4). Hence, the recourse matrix $W(\xi)$ may be random even if the jth recourse matrix W_{jj} for the decision y_j is fixed, but (at least) one of the technology matrices $W_{jj-1}(\xi)$ is random.

Another interesting case appears, second, in risk averse two-stage stochastic programming models, if risk functionals (e.g., the conditional value-at-risk [22]) are incorporated into two-stage stochastic programs. The conditional or *average value-at-risk* (at level $\alpha \in (0, 1]$) may be defined by

$$\begin{aligned} AVaR_\alpha(z) &= \frac{1}{\alpha}\int_0^\alpha VaR_\gamma(z)d\gamma = \inf\left\{r + \frac{1}{\alpha}\mathbb{E}[\max\{0, -r - z\}] : r \in \mathbb{R}\right\} \\ &= \inf\left\{r_1 + \frac{1}{\alpha}\mathbb{E}[r_2^{(2)}] : r_1 \in \mathbb{R}, r_2 \in \mathbb{R}_+ \times \mathbb{R}_+, r_2^{(1)} - r_2^{(2)} = z + r_1\right\}, \end{aligned} \tag{1.3}$$

where z is a real random variable on some probability space. If the average value-at-risk replaces the expectation in a two-stage model with fixed recourse, the latter is of the form

$$\min\{cx + AVaR_\alpha(q(\xi)y) : x \in X, y \in Y, Wy = h(\xi) - T(\xi)x\}. \tag{1.4}$$

Using the two-stage representation (1.3) of $AVaR_\alpha$, the preceding optimization problem is equivalent to (1.1) with

$$f_0(\xi,(x,r_1)) := cx + r_1 + \inf\left\{ \frac{1}{\alpha} r_2^{(2)} : y \in Y, r_2 \geq 0, r_2^{(1)} - r_2^{(2)} = q(\xi)y + r_1, \right.$$
$$\left. Wy = h(\xi) - T(\xi)x \right\},$$

where (x, r_1) is the first-stage decision varying in $X \times \mathbb{R}$. When introducing the second-stage decision (y, r_2), the recourse cost $q_{\text{avar}}(\xi)$, recourse matrix $W_{\text{avar}}(\xi)$, and cone Y_{avar} take on the form

(1.5)
$$q_{\text{avar}}(\xi) = \begin{pmatrix} 0 \\ 0 \\ \alpha^{-1} \end{pmatrix}, \quad W_{\text{avar}}(\xi) = \begin{pmatrix} W & 0 & 0 \\ q(\xi)^\top & -1 & 1 \end{pmatrix}, \quad \text{and} \quad Y_{\text{avar}} = Y \times \mathbb{R}^2_+ .$$

Hence, the recourse matrix gets random if the recourse cost of the original model is random. The same lower diagonal randomness effect appears if general polyhedral convex risk measures are used instead of $AVaR$ (see [7, section 4.1.1]).

In sections 3 and 4 we characterize the local Lipschitz continuity behavior of the functions $\mathcal{F}_\rho$. We also show that the distances $d_{\mathcal{F}_\rho}$ are bounded by Fortet–Mourier (type) metrics and that the metric entropy of $\mathcal{F}_\rho$ in terms of bracketing numbers is reasonably "small." In this way, we obtain new results on stability (Corollaries 3.6 and 4.3 for the cases (i) and (ii), respectively) and on the asymptotic behavior of nonparametric statistical estimates (Theorem 5.2) of random recourse models.

2. Quantitative stability. Given the original probability measure P and a perturbation Q of P we will give quantitative estimates of the distance between $(v(Q), S_\varepsilon(Q))$ and $(v(P), S_\varepsilon(P))$ in terms of a probability metric of the type (1.2). Our analysis will be based on the general perturbation results for optimization models in [23, section 7J].

Let us now introduce functions, spaces, and probability measures that are useful for characterizing classes of probability distributions such that the stochastic program (1.1) is well defined and one can proceed with the perturbation analysis. We consider

$$\mathcal{F} = \{f_0(\cdot, x) \,:\, x \in X\},$$
$$\mathcal{P}_{\mathcal{F}} = \left\{ Q \in \mathcal{P}(\Xi) : \int_\Xi \inf_{x \in X \cap \rho\mathbb{B}} f_0(\xi, x) Q(d\xi) > -\infty \,, \right.$$
$$\left. \sup_{x \in X \cap \rho\mathbb{B}} \int_\Xi f_0(\xi, x) Q(d\xi) < \infty \quad \forall\, \rho > 0 \right\},$$

where $\mathbb{B}$ is the closed unit ball in $\mathbb{R}^m$. We note that the infimum function $\xi \mapsto \inf_{x \in X \cap \rho\mathbb{B}} f_0(\xi, x)$ is measurable for each $\rho > 0$ as f_0 is a random lsc function; cf. [23, Theorem 14.37].

For any $\rho > 0$ and probability measures P, $Q \in \mathcal{P}_{\mathcal{F}}$ we consider their $d_{\mathcal{F},\rho}$-distance defined by

$$d_{\mathcal{F},\rho}(P, Q) = \sup_{x \in X \cap \rho\mathbb{B}} \left| \mathbb{E}^P f_0(x) - \mathbb{E}^Q f_0(x) \right| .$$

Hence, $d_{\mathcal{F},\rho}$ is a distance of type (1.2), where the relevant class of functions is $\mathcal{F}_\rho = \{f_0(\cdot, x) : x \in X \cap \rho\mathbb{B}\}$. It is nonnegative, finite, and symmetric and satisfies the

triangle inequality; i.e., it is a semimetric on $\mathcal{P}_{\mathcal{F}}$. In general, however, the class $\mathcal{F}_\rho$ will not be rich enough to guarantee that $d_{\mathcal{F},\rho}(P,Q)=0$ implies $P=Q$. A valuable consequence of the definition of the class $\mathcal{P}_{\mathcal{F}}$ is that the function $x \mapsto \mathbb{E}^Q f_0(x) = \int_\Xi f_0(\xi,x)\,Q(d\xi)$ is lsc at any Q belonging to $\mathcal{P}_{\mathcal{F}}$ by appealing to Fatou's lemma. Moreover, it is convex on $\mathbb{R}^m$ and finite on X for any such Q.

Since our statements and proofs rely extensively on estimates for the epi-distance between (lsc) functions, we include a brief review of the relevant definitions and implications. Let $d_C(x)=d(x,C)$ denote the distance of a point to a nonempty closed set. The *ρ-distance* between two nonempty closed sets is by definition

$$\mathbb{d}_\rho(C,D) = \sup_{||x||\le\rho} |d_C(x) - d_D(x)|.$$

In fact, it is just a pseudodistance from which one can build a metric on the hyperspace of closed sets, for example, by setting $\mathbb{d}(C,D)=\int_0^\infty \mathbb{d}_\rho(C,D)e^{-\rho}\,d\rho$. Estimates for the ρ-distance can be obtained by relying on a "truncated" Pompeiu–Hausdorff-type distance:

$$\hat{\mathbb{d}}_\rho(C,D) = \inf\{\eta\ge 0:\ C\cap\rho\mathbb{B}\subset D+\eta\mathbb{B},\ \ D\cap\rho\mathbb{B}\subset C+\eta\mathbb{B}\}.$$

Indeed one always has [23, Proposition 4.37(a)]

$$\hat{\mathbb{d}}_\rho(C_1,C_2)\le \mathbb{d}_\rho(C_1,C_2)\le \hat{\mathbb{d}}_{\rho'}(C_1,C_2)$$

for $\rho'\ge 2\rho+\max\{d_{C_1}(0),d_{C_2}(0)\}$. Our main result is stated in terms of this latter distance notion. If we let $\rho\to\infty$, we end up with the Pompeiu–Hausdorff distance

$$\mathbb{d}_\infty(C,D)=\lim_{\rho\to\infty}\mathbb{d}_\rho(C,D)=\lim_{\rho\to\infty}\hat{\mathbb{d}}_\rho(C,D)$$

between the closed nonempty sets C and D; see [23, Corollary 4.38].

The distance between (lsc) functions is measured in terms of the distance between their epi-graphs, so for $\rho>0$,

$$\mathbb{d}_\rho(f,g)=\mathbb{d}_\rho(\operatorname{epi} f,\operatorname{epi} g),\qquad \hat{\mathbb{d}}_\rho(f,g)=\hat{\mathbb{d}}_\rho(\operatorname{epi} f,\operatorname{epi} g),$$

and $\mathbb{d}(f,g)=\mathbb{d}(\operatorname{epi} f,\operatorname{epi} g)$. However, since our sets are epi-graphs (in $\mathbb{R}^{m+1}$), it is convenient to rely on the "unit ball" to be $\mathbb{B}\times[-1,1]$; this brings us to an "auxiliary" distance $\hat{\mathbb{d}}^+_\rho(f_1,f_2)$ defined as the infimum of all $\eta\ge 0$ such that for all $x\in\rho\mathbb{B}$,

$$\min_{y\in\mathbb{B}(x,\eta)} f_2(y)\le\max\{f_1(x),-\rho\}+\eta,\qquad \min_{y\in\mathbb{B}(x,\eta)} f_1(y)\le\max\{f_2(x),-\rho\}+\eta.$$

For lsc $f_1,\ f_2:\mathbb{R}^n\to\mathbb{R}$, not identically ∞, one has [23, Theorem 7.61]

$$\hat{\mathbb{d}}^+_{\rho/\sqrt{2}}(f_1,f_2)\le\hat{\mathbb{d}}_\rho(f_1,f_2)\le\sqrt{2}\,\hat{\mathbb{d}}^+_\rho(f_1,f_2).$$

Our first stability result, already announced in [5], is concerned with the solution set $S(P)$, rather than $S_\varepsilon(P)$, which will be dealt with later.

THEOREM 2.1. *Let $P\in\mathcal{P}_{\mathcal{F}}$, and suppose $S(P)$ is nonempty and bounded. Then there exist constants $\rho>0$ and $\delta>0$ such that*

$$|v(P)-v(Q)|\le d_{\mathcal{F},\rho}(P,Q),$$
$$\emptyset\ne S(Q)\subset S(P)+\Psi_P(d_{\mathcal{F},\rho}(P,Q))\mathbb{B}$$

hold for all $Q \in \mathcal{P}_{\mathcal{F}}$ *with* $d_{\mathcal{F},\rho}(P,Q) < \delta$, *where* Ψ_P *is a* conditioning function *associated with our given problem* (1.1)*; more precisely,*

$$\Psi_P(\eta) := \eta + \psi_P^{-1}(2\eta), \quad \eta \geq 0,$$

with

$$\psi_P(\tau) := \min\left\{\mathbb{E}^P f_0(x) - v(P) : d(x, S(P)) \geq \tau, x \in X\right\}, \quad \tau \geq 0.$$

Proof. For any $Q \in \mathcal{P}_{\mathcal{F}}$, the function $\mathbb{E}^Q f_0$ is lsc, proper, and convex. Define

$$F_Q(x) := \begin{cases} \mathbb{E}^Q f_0(x), & x \in X, \\ +\infty & \text{else} \end{cases}$$

for each $Q \in \mathcal{P}_{\mathcal{F}}$ and rely on [23, Theorem 7.64] to derive the result. Let $\bar{\rho} > 0$ be chosen such that $S(P) \subset \bar{\rho}\mathbb{B}$ and $v(P) \geq -\bar{\rho}$. For $\rho > \bar{\rho}$ and δ such that $0 < \delta < \min\{\frac{1}{2}(\rho - \bar{\rho}), \frac{1}{2}\psi_P(\frac{1}{2}(\rho - \bar{\rho}))\}$, since F_Q and F_P are convex, Theorem 7.64 of [23] yields the estimates

$$\begin{aligned} |v(P) - v(Q)| &\leq \hat{\boldsymbol{d}}_\rho^+(\mathbb{E}^P f_0, \mathbb{E}^Q f_0), \\ \emptyset \neq S(Q) &\subseteq S(P) + \Psi_P(\hat{\boldsymbol{d}}_\rho^+(\mathbb{E}^P f_0, \mathbb{E}^Q f_0))\mathbb{B} \end{aligned}$$

for any $Q \in \mathcal{P}_{\mathcal{F}}$ with $\hat{\boldsymbol{d}}_\rho^+(\mathbb{E}^P f_0, \mathbb{E}^Q f_0) < \delta$.

Now, let η be chosen such that $\eta \geq \max_{x \in X \cap \rho\mathbb{B}} |\mathbb{E}^P f_0(x) - \mathbb{E}^Q f_0(x)|$. Clearly, the inequalities

$$\begin{aligned} \min_{y \in x + \eta\mathbb{B}} F_Q(y) &\leq \max\{F_P(x), -\rho\} + \eta, \\ \min_{y \in x + \eta\mathbb{B}} F_P(y) &\leq \max\{F_Q(x), -\rho\} + \eta \end{aligned}$$

are trivially satisfied when $x \notin X$. When $x \in X \cap \rho\mathbb{B}$, we have

$$\begin{aligned} \min_{y \in x + \eta\mathbb{B}} F_Q(y) \leq F_Q(x) \leq F_P(x) + \eta = \max\{F_P(x), -\rho\} + \eta, \\ \min_{y \in x + \eta\mathbb{B}} F_P(y) \leq F_P(x) \leq F_Q(x) + \eta \leq \max\{F_Q(x), -\rho\} + \eta, \end{aligned}$$

and, thus, $\hat{\boldsymbol{d}}_\rho^+(F_P, F_Q) \leq \eta$. Letting η pass to its lower limit leads to

$$\hat{\boldsymbol{d}}_\rho^+(F_P, F_Q) \leq \max_{x \in X \cap \rho\mathbb{B}} |\mathbb{E}^P f_0(x) - \mathbb{E}^Q f_0(x)| = d_{\mathcal{F},\rho}(P,Q). \tag{2.1}$$

Since the function Ψ_P is increasing, the proof is complete. □

Simple examples of two-stage stochastic programs show that, in general, the set-valued mapping $S(.)$ is not inner semicontinuous at P (cf. [24, Example 26]). Furthermore, explicit descriptions of conditioning functions ψ_P of stochastic programs (like linear or quadratic growth at solution sets) are known only in some specific cases—for example, for linear two-stage stochastic programs with finite discrete distribution or with strictly positive densities of random right-hand sides [28].

As we shall see, we are in much better shape when we consider the stability properties of the sets $S_\varepsilon(\cdot)$ of ε-approximate solutions. Indeed, $S_\varepsilon(\cdot)$ even satisfies a Lipschitz property under rather mild assumptions.

THEOREM 2.2. *Let $P, Q \in \mathcal{P}_{\mathcal{F}}$ and such that the corresponding solution sets $S(P)$ and $S(Q)$ are nonempty. Then there exist constants $\rho > 0$ and $\bar{\varepsilon} > 0$ such that*

$$\hat{\mathbb{d}}_{\rho}(S_{\varepsilon}(P), S_{\varepsilon}(Q)) \leq \frac{4\rho}{\varepsilon} d_{\mathcal{F},\rho+\varepsilon}(P, Q)$$

holds for any $\varepsilon \in (0, \bar{\varepsilon})$, where $d_{\mathcal{F},\rho+\varepsilon}(P, Q) < \varepsilon$.

Proof. The assumptions imply that both $\mathbb{E}^P f_0$ and $\mathbb{E}^Q f_0$ are proper, lsc, and convex on $\mathbb{R}^m$. Let ρ_0 be chosen such that both $S(P) \cap \rho_0 \mathbb{B}$ and $S(Q) \cap \rho_0 \mathbb{B}$ are nonempty and $\min\{v(P), v(Q)\} \geq -\rho_0$. For $\rho > \rho_0$ and $0 < \varepsilon < \bar{\varepsilon} = \rho - \rho_0$, one obtains, from the proof of [23, Theorem 7.69], the inclusion

$$S_{\varepsilon}(P) \cap \rho \mathbb{B} \subseteq S_{\varepsilon}(Q) + \frac{2\eta}{\varepsilon + 2\eta} 2\rho \mathbb{B} \subseteq S_{\varepsilon}(Q) + \frac{4\rho}{\varepsilon} \eta \mathbb{B}$$

for all $\eta > \hat{\mathbb{d}}^+_{\rho+\varepsilon}(\mathbb{E}^P f_0, \mathbb{E}^Q f_0)$. This implies

$$S_{\varepsilon}(P) \cap \rho \mathbb{B} \subseteq S_{\varepsilon}(Q) + \frac{4\rho}{\varepsilon} \hat{\mathbb{d}}^+_{\rho+\varepsilon}(\mathbb{E}^P f_0, \mathbb{E}^Q f_0) \mathbb{B}.$$

The same argument works with P and Q interchanged. Finally, we appeal to the estimate (2.1) to complete the proof. □

The above estimate for ε-approximate solution sets allows for the solution sets to be unbounded and, thus, extends [24, Theorem 13]. The result becomes somewhat more tangible if the original solution set $S(P)$ is assumed to be bounded.

COROLLARY 2.3. *Let $P \in \mathcal{P}_{\mathcal{F}}$ and $S(P)$ be nonempty and bounded. Then there exist constants $\hat{\rho} > 0$ and $\hat{\varepsilon} > 0$ such that*

$$\mathbb{d}_{\infty}(S_{\varepsilon}(P), S_{\varepsilon}(Q)) \leq \frac{4\hat{\rho}}{\varepsilon} d_{\mathcal{F},\hat{\rho}+\varepsilon}(P, Q)$$

holds for any $\varepsilon \in (0, \hat{\varepsilon})$ and $Q \in \mathcal{P}_{\mathcal{F}}$ such that $d_{\mathcal{F},\hat{\rho}+\varepsilon}(P, Q) < \varepsilon$.

Proof. Let δ and ρ be the constants from Theorem 2.1, and put $\hat{\varepsilon} = \delta$. Let $\varepsilon \in (0, \hat{\varepsilon})$ and $Q \in \mathcal{P}_{\mathcal{F}}$ such that $d_{\mathcal{F},\rho+\varepsilon}(P, Q) < \varepsilon$. Then $S(Q)$ is also nonempty and bounded. Since the functions $\mathbb{E}^P f_0$ and $\mathbb{E}^Q f_0$ are lsc and convex, the level sets $S_{\hat{\varepsilon}}(P)$ and $S_{\hat{\varepsilon}}(Q)$ are bounded since the sets $S_0(P)$ and $S_0(Q)$ are bounded (cf. [20, Corollary 8.7.1]). Next we choose ρ_0 as in Theorem 2.2 and $\hat{\rho}$ such that $\hat{\rho} > \max\{\rho, \rho_0 + \hat{\varepsilon}\}$ and both level sets $S_{\hat{\varepsilon}}(P)$ and $S_{\hat{\varepsilon}}(Q)$ are contained in $\hat{\rho}\mathbb{B}$. Then the result follows from Theorem 2.2 by taking into account that

$$\hat{\mathbb{d}}_{\hat{\rho}}(S_{\varepsilon}(P), S_{\varepsilon}(Q)) = \mathbb{d}_{\infty}(S_{\varepsilon}(P), S_{\varepsilon}(Q))$$

holds because of the choice of $\hat{\rho}$. □

The results illuminate the role of the probability distances $d_{\mathcal{F},\rho}$ given that the parameter $\rho > 0$ is properly chosen. These probability metrics process the minimal information about problem (1.1) and allow us to derive remarkable stability properties for the optimal values and (approximate) solutions. Clearly, the preceding stability results remain valid if the set $\mathcal{F}_{\rho}$ is enlarged to a set $\hat{\mathcal{F}}$ and the set $\mathcal{P}_{\mathcal{F}}$ is reduced to a subset on which the new distance $d_{\hat{\mathcal{F}}}$ is finite and well defined.

Hence, it is important to identify classes $\hat{\mathcal{F}}$ of functions that contain $\{f_0(\cdot, x) : x \in X \cap \rho\mathbb{B}\}$ for any $\rho > 0$. For many convex stochastic programming problems the functions $f_0(\cdot, x)$, $x \in X$, are locally Lipschitz continuous on Ξ with certain Lipschitz constants $L(r)$ on the sets $\{\xi \in \Xi : \|\xi - \xi_0\| \leq r\}$ for some $\xi_0 \in \Xi$ and any $r > 0$. In

many cases, the growth modulus $L(r)$ does not depend on x, particularly when x is varying only in a bounded subset of $\mathbb{R}^m$. Hence, function classes of the form

$$\mathcal{F}_H := \{f : \Xi \to \mathbb{R} : f(\xi) - f(\tilde{\xi}) \leq \max\{1, H(\|\xi - \xi_0\|), H(\|\tilde{\xi} - \xi_0\|)\}\|\xi - \tilde{\xi}\| \ \forall \xi, \tilde{\xi} \in \Xi\}$$

are of particular interest, where $H : \mathbb{R}_+ \to \mathbb{R}_+$ is nondecreasing, $H(0) = 0$, and $\xi_0 \in \Xi$. The distances introduced in (1.2), but with $\mathcal{F} = \mathcal{F}_H$, i.e.,

$$d_{\mathcal{F}_H}(P, Q) = \sup\left\{\left|\int_\Xi f(\xi)P(d\xi) - \int_\Xi f(\xi)Q(d\xi)\right| : f \in \mathcal{F}_H\right\},$$

are so-called *Fortet–Mourier metrics*, denoted by ζ_H and defined on

$$\mathcal{P}_H(\Xi) := \left\{Q \in \mathcal{P}(\Xi) : \int_\Xi \max\{1, H(\|\xi - \xi_0\|)\}\|\xi - \xi_0\|Q(d\xi) < \infty\right\} \tag{2.2}$$

(cf. [8, 17]). Important special cases come to light when the function H has the polynomial form $H(t) := t^{r-1}$ for $r \geq 1$. The corresponding function classes and distances are denoted by $\mathcal{F}_r$ and ζ_r, respectively. The distances ζ_r are well defined on the set

$$\mathcal{P}_r(\Xi) := \left\{Q \in \mathcal{P}(\Xi) : \int_\Xi \|\xi\|^r Q(d\xi) < \infty\right\} \tag{2.3}$$

of probability measures having finite rth order moments.

3. Stability of two-stage recourse models. We consider the linear two-stage stochastic program with recourse,

$$\min\left\{cx + \int_\Xi q(\xi)y(\xi)P(d\xi) : W(\xi)y(\xi) = h(\xi) - T(\xi)x,\ y(\xi) \in Y,\ x \in X\right\}, \tag{3.1}$$

where $c \in \mathbb{R}^m$, $X \subseteq \mathbb{R}^m$, and $\Xi \subseteq \mathbb{R}^s$ are polyhedral, $Y \subseteq \mathbb{R}^{\overline{m}}$ is a polyhedral cone, and $P \in \mathcal{P}(\Xi)$. We assume that $q(\xi) \in \mathbb{R}^{\overline{m}}$, $h(\xi) \in \mathbb{R}^d$, the recourse matrix $W(\xi) \in \mathbb{R}^{d \times \overline{m}}$, and the technology matrix $T(\xi) \in \mathbb{R}^{d \times n}$ may depend affinely on $\xi \in \Xi$.

Denoting by $\Phi(\xi, q(\xi), h(\xi) - T(\xi)x)$ the value of the optimal second-stage decision, problem (3.1) may be rewritten equivalently as a minimization problem with respect to the first stage decision x. We define the function $f_0 : \Xi \times \mathbb{R}^m \to \overline{\mathbb{R}}$ by

$$f_0(\xi, x) = \begin{cases} cx + \Phi(\xi, q(\xi), h(\xi) - T(\xi)x) & \text{if } h(\xi) - T(\xi)x \in W(\xi)Y,\ D(\xi) \neq \emptyset, \\ +\infty & \text{otherwise,} \end{cases}$$

where the optimal value function Φ and the dual feasible set $D(\xi)$ are given by

$$\Phi(\xi, u, t) := \inf\{uy : W(\xi)y = t, y \in Y\}, \qquad (\xi, u, t) \in \Xi \times \mathbb{R}^{\overline{m}} \times \mathbb{R}^d,$$
$$D(\xi) := \left\{z \in \mathbb{R}^d : W(\xi)^\top z - q(\xi) \in Y^*\right\}, \quad \xi \in \Xi,$$

with $W(\xi)^\top$ denoting the transpose of $W(\xi)$ and Y^* the polar cone of Y.

The (equivalent) minimization problem can thus be expressed as

$$\min\left\{\int_\Xi f_0(\xi, x)P(d\xi) : x \in X\right\}. \tag{3.2}$$

In order to utilize the general stability results of section 2, we first recall some well-known properties of the function Φ (cf. [34]).

LEMMA 3.1. *For any $\xi \in \Xi$, the function $\Phi(\xi, \cdot, \cdot)$ is finite and continuous on the polyhedral set $\mathcal{D}(\xi) \times W(\xi)Y$, where $\mathcal{D}(\xi) := \{u \in \mathbb{R}^{\overline{m}} : \{z \in \mathbb{R}^d : W(\xi)^\top z - u \in Y^*\} \neq \emptyset\}$. Furthermore, the function $\Phi(\xi, u, \cdot)$ is piecewise linear convex on the polyhedral set $W(\xi)Y$ for fixed $u \in \mathcal{D}(\xi)$, and $\Phi(\xi, \cdot, t)$ is piecewise linear concave on $\mathcal{D}(\xi)$ for fixed $t \in W(\xi)Y$.*

We impose the following conditions on problem (3.2).

(A1) *Relatively complete recourse*: For any $(\xi, x) \in \Xi \times X$, $h(\xi) - T(\xi)x \in W(\xi)Y$.

(A2) *Dual feasibility*: $D(\xi) \neq \emptyset$ holds for all $\xi \in \Xi$.

Conditions (A1) and (A2) are standard and render problem (3.2) well defined. Due to Lemma 3.1 they imply that f_0 is a convex random lsc function with $\Xi \times X \subseteq \operatorname{dom} f_0$. As earlier, with the notation

$$\mathcal{F}_\rho := \{f_0(\cdot, x) : x \in X \cap \rho\mathbb{B}\}, \tag{3.3}$$

we obtain our first stability result for model (3.1) as immediate consequences of Theorem 2.1 and Corollary 2.3.

THEOREM 3.2. *Suppose the stochastic program satisfies the relatively complete recourse* (A1) *and the dual feasibility* (A2) *conditions, $P \in \mathcal{P}_\mathcal{F}$, and $S(P)$ is nonempty and bounded. Then there exist constants $\rho > 0$ and $\hat{\varepsilon} > 0$ such that*

$$|v(P) - v(Q)| \leq d_{\mathcal{F},\rho}(P, Q),$$
$$\mathbb{d}_\infty(S_\varepsilon(P), S_\varepsilon(Q)) \leq \frac{4\rho}{\varepsilon} d_{\mathcal{F},\rho+\varepsilon}(P, Q)$$

hold for any $\varepsilon \in (0, \hat{\varepsilon})$ and each $Q \in \mathcal{P}_\mathcal{F}$ such that $d_{\mathcal{F},\rho+\varepsilon}(P, Q) < \varepsilon$.

The theorem establishes Lipschitz stability of $v(.)$ and S_ε in the two-stage case for fairly general situations. It extends the results in [24, section 3.1] to two-stage models with *random* recourse. However, the set of (perturbed) probability measures $\mathcal{P}_\mathcal{F}$ and, in particular, the metrics $d_{\mathcal{F},\rho}$ are rather sophisticated and could be difficult to use in applications.

To overcome this difficulty, we need to explore quantitative continuity properties of the integrand f_0. Such properties are well known in case of *fixed recourse*, i.e., in case $W(\xi) \equiv W$ [36], and have been used to analyze quantitative stability in [18]. Our first result for random recourse matrices follows the ideas in [37]. There, it is shown that (semi)continuity properties of parametric optimal value functions are consequences of the (semi)continuity of the primal and dual feasibility mapping with respect to the relevant parameters. Next, we verify that a local Lipschitz property of the dual feasible set-valued mapping $\xi \mapsto D(\xi)$ in addition to (A1) implies local Lipschitz continuity of $f_0(\cdot, x)$ with the modulus not depending on having x vary only in a bounded set.

PROPOSITION 3.3. *Suppose the stochastic program satisfies the relatively complete recourse* (A1) *and the dual feasibility* (A2) *conditions. Assume also that the mapping $\xi \mapsto D(\xi)$ is bounded-valued and locally Lipschitz continuous on Ξ with respect to the Pompeiu–Hausdorff distance (on the subsets of $\mathbb{R}^d$); i.e., there exists a constant $L > 0$, an element $\xi_0 \in \Xi$, and a nondecreasing function $h : \mathbb{R}_+ \to \mathbb{R}_+$ with $h(0) = 0$ such that*

$$\mathbb{d}_\infty(D(\xi), D(\tilde{\xi})) \leq L \max\{1, h(\|\xi - \xi_0\|), h(\|\tilde{\xi} - \xi_0\|)\}\|\xi - \tilde{\xi}\| \tag{3.4}$$

holds for all $\xi, \tilde{\xi} \in \Xi$.

Then, for any $\rho > 0$, there exist constants $\hat{L} > 0$ and $\hat{L}(\rho) > 0$ such that

$$f_0(\xi, x) - f_0(\tilde{\xi}, x) \le \hat{L}(\rho) \max\{1, H(\|\xi - \xi_0\|), H(\|\tilde{\xi} - \xi_0\|)\}\|\xi - \tilde{\xi}\|, \tag{3.5}$$

$$f_0(\xi, x) - f_0(\xi, \tilde{x}) \le \hat{L} \max\{1, H(\|\xi - \xi_0\|)\|\xi - \xi_0\|\}\|x - \tilde{x}\| \tag{3.6}$$

for all $\xi, \tilde{\xi} \in \Xi$, $x, \tilde{x} \in X \cap \rho\mathbb{B}$, where H is defined by

$$H(t) := h(t)t \quad \forall t \in \mathbb{R}_+. \tag{3.7}$$

Proof. Let $\rho > 0$. Due to (A1) and (A2), the function $f_0(\cdot, x)$ is real-valued for every $x \in X$. For any $x, \tilde{x} \in X \cap \rho\mathbb{B}$, and $\xi, \tilde{\xi} \in \Xi$, one has the estimate

$$f_0(\xi, x) - f_0(\tilde{\xi}, \tilde{x}) \le cx + (h(\xi) - T(\xi)x)z^*(\xi) - (h(\tilde{\xi}) - c\tilde{x} - T(\tilde{\xi})\tilde{x})z(\tilde{\xi}), \tag{3.8}$$

where $z^*(\xi) \in D(\xi)$ is a dual solution of the second-stage problem and $z(\tilde{\xi})$ is some element in $D(\tilde{\xi})$. We denote by $\bar{z}(\tilde{\xi}; \xi)$ the projection of $z^*(\xi)$ onto $D(\tilde{\xi})$, i.e.,

$$d(z^*(\xi), D(\tilde{\xi})) = \|z^*(\xi) - \bar{z}(\tilde{\xi}; \xi)\|,$$

yielding

$$\|z^*(\xi) - \bar{z}(\tilde{\xi}; \xi)\| \le \boldsymbol{d}_\infty(D(\xi), D(\tilde{\xi})) \le L \max\{1, h(\|\xi - \xi_0\|), h(\|\tilde{\xi} - \xi_0\|)\}\|\xi - \tilde{\xi}\|. \tag{3.9}$$

As $D(\xi_0)$ is bounded, there exists $r > 0$ such that $\|z\| \le r$ for each $z \in D(\xi_0)$. As the estimate

$$d(\bar{z}(\tilde{\xi}; \xi), D(\xi_0)) \le L \max\{1, h(\|\tilde{\xi} - \xi_0\|)\}\|\tilde{\xi} - \xi_0\|$$

holds for all $\xi, \tilde{\xi} \in \Xi$, according to (3.4), we have

$$\|\bar{z}(\tilde{\xi}; \xi)\| \le \max\{r, L\} \max\{1, h(\|\tilde{\xi} - \xi_0\|)\}\|\tilde{\xi} - \xi_0\|. \tag{3.10}$$

Now, we proceed with our estimate (3.8) when $x = \tilde{x}$, exploiting the affine linearity of $h(\cdot)$ and $T(\cdot)$, (3.9) and (3.10). Setting $z(\tilde{\xi}) := \bar{z}(\tilde{\xi}; \xi)$ we obtain

$$\begin{aligned}
f_0(\xi, x) - f_0(\tilde{\xi}, x) &\le (h(\xi) - T(\xi)x)(z^*(\xi) - \bar{z}(\tilde{\xi}; \xi)) \\
&\quad - ((h(\tilde{\xi}) - h(\xi)) - (T(\tilde{\xi}) - T(\xi))x)\bar{z}(\tilde{\xi}; \xi) \\
&\le \|h(\xi) - T(\xi)x\|\|z^*(\xi) - \bar{z}(\tilde{\xi}; \xi)\| \\
&\quad + (\|h(\tilde{\xi}) - h(\xi)\| + \|T(\tilde{\xi}) - T(\xi)\|\|x\|)\|\bar{z}(\tilde{\xi}; \xi)\| \\
&\le \Big(KL(1 + \rho) \max\{1, \|\xi - \xi_0\|\} \max\{1, h(\|\xi - \xi_0\|), h(\|\tilde{\xi} - \xi_0\|)\} \\
&\quad + \tilde{K} \max\{r, L\}(1 + \rho) \max\{1, h(\|\tilde{\xi} - \xi_0\|)\}\|\tilde{\xi} - \xi_0\|\Big)\|\xi - \tilde{\xi}\| \\
&\le \bar{L}(1 + \rho) \max\{1, H(\|\xi - \xi_0\|), H(\|\tilde{\xi} - \xi_0\|)\}\|\xi - \tilde{\xi}\|
\end{aligned}$$

for each $\xi, \tilde{\xi} \in \Xi$, and some positive constants K, $\tilde{K}$, and $\bar{L}$. Thus, (3.5) is proved with $\hat{L}(\rho) = \bar{L}(1 + \rho)$. Finally, we return to (3.8) in case $\xi = \tilde{\xi}$; choosing $\bar{z}(\xi) = z^*(\xi)$, we arrive at the estimate

$$\begin{aligned}
f_0(\xi, x) - f_0(\xi, \tilde{x}) &\le c(x - \tilde{x}) + T(\xi)(\tilde{x} - x)z^*(\xi) \le (\|c\| + \|T(\xi)\|\|z^*(\xi)\|)\|x - \tilde{x}\| \\
&\le \hat{L} \max\{1, H(\|\xi - \xi_0\|)\|\xi - \xi_0\|\}\|x - \tilde{x}\|
\end{aligned}$$

for some constant $\hat{L} > 0$ and all $\xi \in \Xi$, $x, \tilde{x} \in X \cap \rho\mathbb{B}$. Here, we used that $\|z^*(\xi)\|$ can be bounded in the same way as $\bar{z}(\tilde{\xi}; \xi)$ in (3.10). □

The next examples illustrate the local Lipschitz continuity property (3.4) of the dual feasibility mapping D.

Example 3.4. Let $\overline{m} = 4$, $d = 2$, $Y = \mathbb{R}^4_+$, and $\Xi = \mathbb{R}$, and consider the random recourse costs and matrix

$$W(\xi) = \begin{pmatrix} 1 & -1 & 0 & 0 \\ -\xi & 0 & 1 & -1 \end{pmatrix}, \qquad q(\xi) = \begin{pmatrix} 0 \\ 0 \\ \xi \\ -\xi \end{pmatrix}.$$

Then $W(\xi)Y = \mathbb{R}^2$ (complete recourse) and $D(\xi) = [0, \xi^2] \times \{\xi\}$. Hence, the conditions (A1) and (A2) and (3.4) are satisfied with $h(t) \equiv t$.

Example 3.5. We consider the second-stage program arising in the equivalent optimization problem to AVaR minimization (1.4) in section 1. Its dual feasible set is of the form

$$\begin{aligned} D_{\mathrm{avar}}(\xi) &= \{z = (z_1, z_2) \in \mathbb{R}^d \times \mathbb{R} : W_{\mathrm{avar}}(\xi)^\top z - q_{\mathrm{avar}}(\xi) \in Y^*_{\mathrm{avar}}\} \\ &= \{(z_1, z_2) \in \mathbb{R}^d \times [0, \alpha^{-1}] : W^\top z_1 + q(\xi) z_2 \in Y^*\} \\ &= \{(z_1, u) \in \mathbb{R}^d \times \mathbb{R}^{\overline{m}} : W^\top z_1 + u \in Y^*, u \in [0, \alpha^{-1}] q(\xi)\} \end{aligned}$$

due to (1.5), where $u \in [0, \alpha^{-1}]q(\xi)$ means that, for every $j = 1, \ldots, \overline{m}$, $0 \le u_j \le \alpha^{-1} q_j(\xi)$ holds if $q_j(\xi) \ge 0$ and $\alpha^{-1} q_j(\xi) \le u_j \le 0$ otherwise. Hence, if (A2) is satisfied, the set-valued mapping $\xi \to D_{\mathrm{avar}}(\xi)$ is Lipschitz continuous on Ξ with respect to the Pompeiu–Hausdorff distance $\boldsymbol{d}_\infty$ since its graph is convex polyhedral [35]. This means that Proposition 3.3 applies with $h(t) \equiv 1$.

We can reformulate the conclusions of the preceding proposition in terms of the Fortet–Mourier metrics defined on $\mathcal{P}_H(\Xi)$, the space (2.2) of probability measures.

COROLLARY 3.6. *Let the assumptions of Proposition* 3.3 *be satisfied,* $P \in \mathcal{P}_H(\Xi)$*, and* $S(P)$ *be nonempty and bounded. Then there exist constants* $\hat{L} > 0$*,* $\rho > 0$*, and* $\hat{\varepsilon} > 0$ *such that*

$$|v(P) - v(Q)| \le \hat{L} \zeta_H(P, Q),$$

$$\boldsymbol{d}_\infty(S_\varepsilon(P), S_\varepsilon(Q)) \le \frac{4\rho\hat{L}}{\varepsilon} \zeta_H(P, Q)$$

hold for any $\varepsilon \in (0, \hat{\varepsilon})$ *and each* $Q \in \mathcal{P}_H(\Xi)$ *such that* $\zeta_H(P, Q) < \varepsilon$*, where* H *is defined by* (3.7)*, and* $\zeta_H(P, Q)$ *is the Fortet–Mourier metric on* $\mathcal{P}_H(\Xi)$*.*

Proof. The estimate (3.5) implies $d_{\mathcal{F},\rho}(P, Q) \le \hat{L} \zeta_H(P, Q)$ with $\hat{L} = \hat{L}(\rho)$, and, hence, the result follows from Theorem 3.2. □

When $W(\xi) \equiv W$, the mapping $\xi \mapsto D(\xi)$ is even Lipschitz continuous with respect to the Pompeiu–Hausdorff distance $\boldsymbol{d}_\infty$ due to [35]. Hence, $H(t) \equiv t$ and $\mathcal{F}_H = \mathcal{F}_2$, and then the previous result boils down to [18, Proposition 3.2].

4. Two-stage multiperiod models. If the second stage of a stochastic program with recourse models a (stochastic) dynamical decision process (see section 1), our two-stage problem takes on the form

(4.1)

$$\min\left\{ cy_0 + \sum_{j=1}^{l} q_j(\xi) y_j : y_0 \in X, y_j \in Y_j, W_{jj} y_j = h_j(\xi) - W_{jj-1}(\xi) y_{j-1}, j = 1, \ldots, l \right\},$$

where for $j = 1, \dots, l$, $Y_j \in \mathbb{R}^{\overline{m}_j}$ are polyhedral sets for some finite l and first-stage decision $x := y_0$; the matrices $W_{j,j-1}(\xi)$ are (potentially) stochastic. Then the second-stage program has separable block structure; i.e., the recourse variable y has the form $y = (y_1, \dots, y_l)$, the polyhedral set Y is the Cartesian product of polyhedral sets $Y_j \in \mathbb{R}^{\overline{m}_j}$, $j = 1, \dots, l$, the element $T(\xi)x$ has the components $T_1(\xi)x := W_{10}(\xi)x$ and $T_j(\xi)x = 0$, $j = 2, \dots, l$, and the random recourse matrix $W(\xi)$ is of the form

$$W(\xi) = \begin{pmatrix} W_{11} & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ W_{21}(\xi) & W_{22} & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & W_{32}(\xi) & W_{33} & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ 0 & 0 & 0 & 0 & \cdots & W_{l-1l-2}(\xi) & W_{l-1l-1} & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & W_{ll-1}(\xi) & W_{ll} \end{pmatrix}, \tag{4.2}$$

i.e., all matrices W_{jj}, $j = 1, \dots, l$, in the diagonal of $W(\xi)$ are nonstochastic. Denoting by $q_j(\xi)$ and $h_j(\xi)$ the components of $q(\xi)$ and $h(\xi)$, respectively, the integrand f_0 is of the form

$$\begin{aligned} f_0(\xi, x) &= cx + \inf \left\{ \sum_{j=1}^{l} q_j(\xi) y_j : W_{jj} y_j = h_j(\xi) - W_{jj-1}(\xi) y_{j-1}, y_j \in Y_j, j = 1, \dots, l \right\} \\ &=: cx + \Psi_1(\xi, x), \end{aligned}$$

where the function Ψ_1 is given by the recursion

$$\Phi_j(\xi, u_{j-1}) := \inf \{ q_j(\xi) y_j + \Psi_{j+1}(\xi, y_j) : W_{jj} y_j = u_{j-1}, y_j \in Y_j \}, \tag{4.3}$$

$$\Psi_j(\xi, y_{j-1}) := \Phi_j(\xi, h_j(\xi) - W_{jj-1}(\xi) y_{j-1}) \tag{4.4}$$

for $j = l, \dots, 1$, where $y_0 = x$ and $\Psi_{l+1}(\xi, y_l) \equiv 0$.

While the continuity and growth properties of the function $f_0(\cdot, x)$ in case $l = 1$ may be derived from Lemma 3.1, we need an extended result for establishing Lipschitz continuity properties of the inf-projection Φ_j for $j = 1, \dots, l$. The results in [39] were developed precisely to deal with the present situation. To state the result, we denote by D^∞ the recession cone of a convex set $D \subseteq \mathbb{R}^m$. It consists of all elements $x_d \in \mathbb{R}^m$ such that $x + \lambda x_d \in D$ for all $x \in D$ and $\lambda \in \mathbb{R}_+$. Clearly, we have $D^\infty = \{0\}$ if D is bounded. Furthermore, D^∞ is polyhedral if D is polyhedral. Next we record [39, Proposition 4.4] and provide a self-contained proof for the convenience of the reader.

LEMMA 4.1. *Let $h \in \mathbb{R}^d$, $W \in \mathbb{R}^{d \times n}$, and $Y \subseteq \mathbb{R}^n$ be polyhedral. Let $u = (u_1, u_2) \in \mathbb{R}^n \times \mathbb{R}^d$ and*

$$\Phi(u) := \inf\{ f(u_1, y) : Wy = h - u_2, y \in Y \}.$$

Assume that $\ker(W) \cap Y^\infty = \{0\}$ *and that f is Lipschitz continuous on $\{(u_1, y) \in \mathbb{R}^n \times Y : \|u_1\| \le r, \|y\| \le r\}$ with constant $L(r)$ for every $r > 0$. Then, $\Phi(\cdot)$ is Lipschitz continuous on $\{(u_1, u_2) \in \operatorname{dom} \Phi : \|u_1\| \le r, \|u_2\| \le r\}$ with constant $L_M L(K_M \max\{1, r\})$ for every $r > 0$, where $L_M \ge 1$ and $K_M \ge 1$ are constants depending only on the set-valued mapping $M(u_2) := \{y \in Y : Wy = h - u_2\}$ from $\mathbb{R}^d$ to $\mathbb{R}^n$.*

Proof. The condition $\ker(W) \cap Y^\infty = \{0\}$ is equivalent to the local boundedness of the mapping M. M is Lipschitz continuous with respect to the Pompeiu–Hausdorff distance d_∞ (with constant $L_M \ge 1$) since its graph is polyhedral [23, Example

9.35]. Since the set $M(u_2)$ is compact, Φ is finite for all pairs (u_1, u_2) such that $u_2 \in \operatorname{dom} M$. Now, let $r > 0$ and $u = (u_1, u_2), \tilde{u} = (\tilde{u}_1, \tilde{u}_2) \in \operatorname{dom}\Phi \cap \{(u_1, u_2) \in \mathbb{R}^n \times \mathbb{R}^d : \|u_1\| \le r, \|u_2\| \le r\}$. Then there exist $y(u_2) \in M(u_2)$ and $y(\tilde{u}_2) \in M(\tilde{u}_2)$ such that $\Phi(u) = f(u_1, y(u_2))$ and $\|y(u_2) - y(\tilde{u}_2)\| \le L_M \|u_2 - \tilde{u}_2\|$. In particular, there exists a constant $K_M \ge 1$ such that

$$\max\{\|y(u_2)\|, \|y(\tilde{u}_2)\|\} \le K_M \max\{1, \|u_2\|, \|\tilde{u}_2\|\} \le K_M \max\{1, r\}.$$

We obtain

$$\begin{aligned}\Phi(\tilde{u}) - \Phi(u) &\le f(\tilde{u}_1, y(\tilde{u}_2)) - f(u_1, y(u_2)) \\ &\le L(K_M \max\{1, r\})(\|\tilde{u}_1 - u_1\| + \|y(\tilde{u}_2) - y(u_2)\|) \\ &\le L_M L(K_M \max\{1, r\})(\|\tilde{u}_1 - u_1\| + \|\tilde{u}_2 - u_2\|),\end{aligned}$$

and that completes the proof. □

PROPOSITION 4.2. *Let $W(\xi)$ be as described by* (4.2). *Assume the relatively complete recourse condition* (A1) *is satisfied and that* $\ker(W_{jj}) \cap Y_j^\infty = \{0\}$ *for* $j = 1, \dots, l-1$. *Then, there exist constants $L > 0$, $\hat{L} > 0$, and $K > 0$ such that the following holds for all $\xi, \tilde{\xi} \in \Xi$ and $x, \tilde{x} \in X \cap \rho\mathbb{B}$:*

$$\begin{aligned}|f_0(\xi, x) - f_0(\tilde{\xi}, x)| &\le L \max\{1, \rho, \|\xi\|^l, \|\tilde{\xi}\|^l\}\|\xi - \tilde{\xi}\|, \\ |f_0(\xi, x) - f_0(\xi, \tilde{x})| &\le \hat{L} \max\{1, \|\xi\|^{l+1}\}\|x - \tilde{x}\|, \\ |f_0(\xi, x)| &\le K \max\{1, \rho, \|\xi\|^{l+1}\}.\end{aligned}$$

Proof. Due to the assumptions, all sets of the form $M_j(v_j) := \{y_j \in Y_j : W_{jj} y_j = v_j\}$ are bounded polyhedra for all $v_j \in \mathbb{R}^{r_j}$ and $j = 1, \dots, l$. Furthermore, the set-valued mappings M_j from $\mathbb{R}^{r_j}$ to $\mathbb{R}^{\overline{m}_j}$ are Lipschitz continuous on $\operatorname{dom} M_j$ with constant L_j. Due to (A1), we have recursively $h_j(\xi) - W_{jj-1}(\xi) y_{j-1} \in \operatorname{dom} M_j$ for all $y_{j-1} \in Y_{j-1}$, $y_0 = x \in X$, $\xi \in \Xi$, and $j = 2, \dots, l$. Hence, if Lemma 4.1 is used recursively by setting $\Phi = \Phi_j$, $f_j(u_1, y_j) := q_j(\xi) y_j + \Psi_{j+1}(\xi, y_j)$ with $u_1 = \xi$ and $u_2 = u_{j-1}$, each subproblem (4.3) is solvable. First we consider the functions Φ_l and Ψ_l:

$$\begin{aligned}\Phi_l(\xi, u_{l-1}) &= \inf\{q_l(\xi) y_l : W_{ll} y_l = u_{l-1}, y_l \in Y_l\}, \\ \Psi_l(\xi, y_{l-1}) &= \Phi_l(\xi, h_l(\xi) - W_{ll-1}(\xi) y_{l-1}).\end{aligned}$$

Then the Lipschitz constant of f_j on $\{(\xi, y_l) \in \Xi \times Y_l : \|\xi\| \le r, \|y_l\| \le r\}$ has the form $L_l \max\{1, r\}$ and Lemma 4.1 implies that Φ_l has the Lipschitz constant $\hat{L}_l \max\{1, r\}$ on $\{(\xi, u_{l-1}) \in \Xi \times \operatorname{dom} M_l : \|\xi\| \le r, \|u_{l-1}\| \le r\}$. Due to the term $W_{ll-1}(\xi) y_{l-1}$ in the definition of Ψ_l, however, the function Ψ_l has the Lipschitz constant $\tilde{L}_l \max\{1, r^2\}$ on $\{(\xi, y_{l-1}) \in \Xi \times Y_{l-1} : \|\xi\| \le r, \|y_{l-1})\| \le r\}$. Since Ψ_l enters the definition of f_{l-1} and the infimum, Φ_{l-1} is Lipschitz continuous with constant $\hat{L}_{l-1} \max\{1, r^2\}$ on $\{(\xi, u_{l-2}) \in \Xi \times \operatorname{dom} M_{l-1} : \|\xi\| \le r, \|u_{l-2}\| \le r\}$ according to Lemma 4.1. Due to the term $W_{l-1l-2}(\xi) y_{l-2}$, the function Ψ_{l-1} is Lipschitz continuous with constant $\tilde{L}_{l-1} \max\{1, r^3\}$ on $\{(\xi, y_{l-2}) \in \Xi \times Y_{l-2} : \|\xi\| \le r, \|y_{l-2})\| \le r\}$, etc. This process may be continued until one concludes that Φ_1 is Lipschitz continuous with constant $\hat{L}_1 \max\{1, r^l\}$ on $\{(\xi, u_0) \in \Xi \times \operatorname{dom} M_1 : \|\xi\| \le r, \|u_0\| \le r\}$. Hence, the function Ψ_1 depending on (ξ, x) satisfies the Lipschitz continuity property

$$|\Psi_1(\xi, x) - \Psi_1(\tilde{\xi}, \tilde{x})| \le \tilde{L}_1 \max\{1, \rho, r^l\}(\max\{1, \rho\}\|\xi - \tilde{\xi}\| + \max\{1, r\}\|x - \tilde{x}\|)$$

on the set $\{(\xi, x) \in \Xi \times X : \|\xi\| \le r, \|x\| \le \rho\}$.

This yields the assertions about f_0 and completes the proof. □

Due to the previous result we obtain

$$\mathcal{P}_{\mathcal{F}} \supseteq \mathcal{P}_{l+1}(\Xi) = \{Q \in \mathcal{P}(\Xi) : \int_{\Xi} \|\xi\|^{l+1} Q(d\xi) < \infty\}$$
$$\text{and} \quad \frac{1}{L \max\{1, \rho\}} f_0(x, \cdot) \in \mathcal{F}_{l+1}(\Xi)$$

for each $x \in X \cap \rho\mathbb{B}$, and arrive, after specializing Theorem 3.2, at the following.

COROLLARY 4.3. *Let $W(\xi)$ be as described by* (4.2). *Assume the relatively complete recourse condition* (A1) *is satisfied and that* $\ker(W_{jj}) \cap Y_j^{\infty} = \{0\}$ *for* $j = 1, \dots, l-1$.

Then there exist constants $L > 0$ and $\hat{\varepsilon} > 0$ such that for any $\varepsilon \in (0, \hat{\varepsilon})$ the estimates

$$|v(P) - v(Q)| \le L\, \zeta_{l+1}(P, Q),$$
$$d_{\infty}(S_{\varepsilon}(P), S_{\varepsilon}(Q)) \le \frac{L}{\varepsilon} \zeta_{l+1}(P, Q)$$

hold whenever $Q \in \mathcal{P}_{l+1}(\Xi)$ and $\zeta_{l+1}(P, Q) < \varepsilon$.

The case $l = 1$ corresponds to the situation of two-stage models with fixed recourse, and that situation was already covered by [24, Theorem 24]. We note that the corollary remains valid for the slightly more general situation that $W_{jj-1}(\xi) y_{j-1}$ in (4.1) is replaced by $\sum_{i=1}^{j-1} W_{ji}(\xi) y_i$, and, hence, all lower diagonal blocks of $W(\xi)$ are random. We also note that the corollary applies to recourse matrices of the form (1.5) in risk averse two-stage models with polyhedral convex risk functionals.

If the recent stability result [10, Theorem 2.1] for linear multistage models is restricted to the two-stage model (4.1), it implies the existence of positive constants L and δ such that

$$|v(P) - v(Q)| \le L\, \ell_{l+1}(P, Q) \tag{4.5}$$

holds for every $Q \in \mathcal{P}_{l+1}(\Xi)$ with $\ell_{l+1}(P, Q) < \delta$; the distance ℓ_r denotes the L_r-*minimal* or Wasserstein metric

$$\ell_r(P, Q) := \left(\inf \left\{ \int_{\Xi \times \Xi} \|\xi - \tilde{\xi}\|^r \eta(d\xi, d\tilde{\xi}) : \eta \in \mathcal{P}(\Xi \times \Xi),\ \pi_1 \eta = P,\ \pi_2 \eta = Q \right\} \right)^{1/r} \tag{4.6}$$

on $\mathcal{P}_r(\Xi)$ for any $r \ge 1$, where π_1 and π_2 denote the projections onto the first and second components, respectively. It is known that sequences in $\mathcal{P}_r(\Xi)$ converge with respect to both metrics ζ_r and ℓ_r if they converge weakly and if their rth order absolute moments converge. To derive a quantitative estimate, let $\eta^* \in \mathcal{P}(\Xi \times \Xi)$ be a solution of the minimization problem on the right-hand side of (4.6). Such solutions exist according to [17, Theorem 8.1.1]. Then the duality theorem [17, Theorem 5.3.2]

for the Fortet–Mourier metric of order r implies, via Hölder's inequality, the estimate

$$\begin{aligned}\zeta_r(P,Q) &\le \int_{\Xi\times\Xi} \max\{1,\|\xi\|,\|\tilde{\xi}\|\}^{r-1}\|\xi-\tilde{\xi}\|\eta^*(d\xi,d\tilde{\xi})\\ &\le \Big(\int_{\Xi\times\Xi}\max\{1,\|\xi\|,\|\tilde{\xi}\|\}^r\eta^*(d\xi,d\tilde{\xi})\Big)^{\frac{r-1}{r}}\Big(\int_{\Xi\times\Xi}\|\xi-\tilde{\xi}\|^r\eta^*(d\xi,d\tilde{\xi})\Big)^{\frac{1}{r}}\\ &= \Big(\int_{\Xi\times\Xi}\max\{1,\|\xi\|,\|\tilde{\xi}\|\}^r\eta^*(d\xi,d\tilde{\xi})\Big)^{\frac{r-1}{r}}\ell_r(P,Q)\\ &\le \Big(1+\int_{\Xi}\|\xi\|^r(P+Q)(d\xi)\Big)^{\frac{r-1}{r}}\ell_r(P,Q).\end{aligned}$$

Since the convergence of probability measures with respect to ℓ_r and ζ_r implies the convergence of their rth order absolute moments, the stability result for optimal values obtained in Corollary 4.3 implies (4.5) (with some constant $L>0$). However, the convergence of $\zeta_r(P,P_n)$ to 0 may be faster than $\ell_r(P,P_n)$ for some sequence (P_n) of probability measures, as illustrated in [18, Example 3.4]. Hence, the stability result for optimal values in Corollary 4.3 strictly extends the estimate (4.5) for multiperiod two-stage stochastic programs.

5. Empirical approximations of two-stage models. Let $\xi_1,\xi_2,\dots,\xi_n,\dots$ be independent and identically distributed Ξ-valued random variables on some probability space $(\Omega,\mathcal{A},\mathbb{P})$ having the common distribution P, i.e., $P=\mathbb{P}\xi_1^{-1}$. We consider the empirical measures

$$P_n(\omega) := \frac{1}{n}\sum_{i=1}^n \delta_{\xi_i(\omega)} \qquad (\omega\in\Omega;\ n\in\mathbb{N})$$

and the *empirical approximation* of the stochastic program (1.1) with sample size n, i.e.,

$$\min\left\{\frac{1}{n}\sum_{i=1}^n f_0(\xi_i(\cdot),x) : x\in X\right\}. \tag{5.1}$$

Since the objective function of (5.1) is a random lsc function from $\mathbb{R}^m\times\Omega$ to $\overline{\mathbb{R}}$, the optimal value $v(P_n(\cdot))$ of (5.1) is measurable from Ω to $\overline{\mathbb{R}}$ and the ε-approximate solution set $S_\varepsilon(P_n(\cdot))$ is a closed-valued measurable set-valued mapping from Ω to $\mathbb{R}^m$ (see Chapter 14 and, in particular, Theorem 14.37 of [23]).

Qualitative and quantitative results on the asymptotic behavior of solutions to (5.1) are given, e.g., in [2, 6, 13] and [12, 15, 16, 18, 30], respectively.

Due to the results in the previous sections, the asymptotic behavior of $v(P_n(\cdot))$ and $S_\varepsilon(P_n(\cdot))$ is closely related to uniform convergence properties of the empirical process

$$\left\{\sqrt{n}(P_n(\cdot)-P)f = \frac{1}{\sqrt{n}}\sum_{i=1}^n (f(\xi_i(\cdot))-Pf)\right\}_{f\in\mathcal{F}}$$

indexed by the class $\mathcal{F}=\{f_0(x,\cdot):x\in X\}$. Here, we set $Qf:=\int_\Xi f(\xi)Q(d\xi)$ for any $Q\in\mathcal{P}(\Xi)$ and $f\in\mathcal{F}$. Uniform convergence properties refer to the convergence, or to the convergence rate, of

$$d_{\mathcal{F}}(P_n(\cdot),P) = \sup_{f\in\mathcal{F}}|P_n(\cdot)f-Pf| \tag{5.2}$$

to 0 in terms of some stochastic convergence. Since the supremum in (5.2) is nonmeasurable in general, the outer probability $\mathbb{P}^*$ (defined by $\mathbb{P}^*(B) = \inf\{\mathbb{P}(A) : B \subset A, A \in \mathcal{A}\}$ for any subset B of Ω) is used to describe convergence in probability and almost surely, respectively (cf. [32]).

The class $\mathcal{F}$ is called a *P–Glivenko–Cantelli class* if the sequence $(d_{\mathcal{F}}(P_n(\cdot), P))$ of random variables converges to 0 $\mathbb{P}^*$-almost surely or, equivalently, in outer probability. The empirical process is called *uniformly bounded in outer probability with tail $C_{\mathcal{F}}(\cdot)$* if the function $C_{\mathcal{F}}(\cdot)$ is defined on $(0, \infty)$ and decreasing to 0, and the estimate

$$\mathbb{P}^*(\{\omega : \sqrt{n}\, d_{\mathcal{F}}(P_n(\omega), P) \geq \varepsilon\}) \leq C_{\mathcal{F}}(\varepsilon) \tag{5.3}$$

holds for all $\varepsilon > 0$ and $n \in \mathbb{N}$.

Whether a given class $\mathcal{F}$ is a P–Glivenko–Cantelli class or the empirical process is uniformly bounded in outer probability depends on the size of the class $\mathcal{F}$ measured in terms of *bracketing numbers*, or of the corresponding *metric entropy numbers* defined as their logarithms (see [32]). To introduce this concept, let $\mathcal{F}$ be a subset of the normed linear space $L_p(\Xi, P)$ (for some $p \geq 1$) equipped with the usual norm $\|f\|_{P,\,p} = (P|f|^p)^{\frac{1}{p}}$. The bracketing number $N_{[\,]}(\varepsilon, \mathcal{F}, L_p(\Xi, P))$ is the minimal number of *brackets* $[l, u] = \{f \in L_p(\Xi, P) : l \leq f \leq u\}$ with $\|l - u\|_{P,\,p} < \varepsilon$ needed to cover $\mathcal{F}$. The following result provides criteria for the desired properties in terms of bracketing numbers. For its proof we refer to [32, Theorem 2.4.1] and [31, Theorem 1.3].

THEOREM 5.1. *Let $\mathcal{F}$ be a class of real-valued functions on Ξ. If*

$$N_{[\,]}(\varepsilon, \mathcal{F}, L_1(\Xi, P)) < \infty \tag{5.4}$$

holds for every $\varepsilon > 0$, then $\mathcal{F}$ is a P–Glivenko–Cantelli class.

If $\mathcal{F}$ is uniformly bounded and there exist constants $r \geq 1$ and $R \geq 1$ such that

$$N_{[\,]}(\varepsilon, \mathcal{F}, L_2(\Xi, P)) \leq \left(\frac{R}{\varepsilon}\right)^r \tag{5.5}$$

for every $\varepsilon > 0$, then the empirical process indexed by $\mathcal{F}$ is uniformly bounded in outer probability with exponential tail $C_{\mathcal{F}}(\varepsilon) = (K(R)\varepsilon r^{-\frac{1}{2}})^r \exp(-2\varepsilon^2)$ with some constant $K(R)$ depending only on R.

Next we consider the class $\mathcal{F} := \mathcal{F}_\rho$ of integrands defined by (3.3) in section 3 and derive conditions implying the assumptions of Theorem 5.1, particularly the assumptions (5.4) and (5.5) for the bracketing numbers $N_{[\,]}(\varepsilon, \mathcal{F}_\rho, L_p(\Xi, P))$ with $p \in \{1, 2\}$.

THEOREM 5.2. *Let the assumptions of Proposition 3.3 be satisfied and $H : \mathbb{R}_+ \to \mathbb{R}_+$ be defined by (3.7). If $P \in \mathcal{P}_H(\Xi)$, then $\mathcal{F}_\rho = \{f_0(\cdot, x) : x \in X \cap \rho\mathbb{B}\}$ is a P-Glivenko–Cantelli class for any $\rho > 0$, i.e.,*

$$\lim_{n\to\infty} \sup_{x \in X \cap \rho\mathbb{B}} \left| \int_\Xi f_0(\xi, x) P_n(\omega)(d\xi) - \int_\Xi f_0(\xi, x) P(d\xi) \right| = 0 \quad \mathbb{P}\text{- a.s.} \tag{5.6}$$

If, in addition, Ξ is bounded, then the empirical process indexed by $\mathcal{F}_\rho$ is uniformly bounded in probability with exponential tail; i.e.,

$$\mathbb{P}\left(\left\{\omega : \sqrt{n} \sup_{x \in X \cap \rho\mathbb{B}} \left| \int_\Xi f_0(\xi, x)(P_n(\omega) - P)(d\xi) \right| \geq \varepsilon \right\}\right) \leq (K(R)\varepsilon r^{-\frac{1}{2}})^r \exp(-2\varepsilon^2) \tag{5.7}$$

holds for some constant $K(R) > 0$, any $\varepsilon > 0$, and $n \in \mathbb{N}$.

Proof. According to (3.6) in Proposition 3.3, the functions $f_0(\xi, \cdot)$ satisfy the Lipschitz property

$$f_0(\xi, x) - f_0(\xi, \tilde{x}) \leq \hat{L} \max\{1, H(\|\xi - \xi_0\|)\|\xi - \xi_0\|\}\|x - \tilde{x}\|$$

for all $x, \tilde{x} \in X \cap \rho\mathbb{B}$, and $\xi \in \Xi$. Setting $F(\xi) := \hat{L} \max\{1, H(\|\xi - \xi_0\|)\|\xi - \xi_0\|\}$ for all $\xi \in \Xi$, we conclude from [32, Theorem 2.7.11] that

$$N_{[]}(2\varepsilon\|F\|_{P,1}, \mathcal{F}_\rho, L_1(\Xi, P)) \leq N(\varepsilon, X \cap \rho\mathbb{B}, \mathbb{R}^m) \leq K\varepsilon^{-m} \tag{5.8}$$

holds for some $K > 0$ and all $\varepsilon > 0$. Since $\|F\|_{P,1}$ is finite, we may replace ε by $\varepsilon/2\|F\|_{P,1}$ in (5.8) and obtain that $N_{[]}(\varepsilon, \mathcal{F}_\rho, L_1(\Xi, P))$ is finite for all $\varepsilon > 0$. Thus, condition (5.4) in Theorem 5.1 is satisfied.

If Ξ is bounded, the class $\mathcal{F}_\rho$ is uniformly bounded and condition (5.5) in Theorem 5.1 is also satisfied due to (5.8). It remains to note that the supremum $\sup_{x \in X \cap \rho\mathbb{B}}$ may be replaced by a supremum with respect to a countable dense subset of $X \cap \rho\mathbb{B}$. Hence, the suprema in (5.6) and (5.7) are measurable with respect to $\mathcal{A}$ and, thus, the outer probability $\mathbb{P}^*$ can be replaced by $\mathbb{P}$. □

When combining the previous result with Theorem 3.2, we arrive at conditions implying a Glivenko–Cantelli result and a large deviation result for the distances of empirical ε-approximate solution sets $S_\varepsilon(P_n(\cdot))$ to $S_\varepsilon(P)$ in the case of the two-stage model (3.2) with random recourse.

6. Conclusions. The quantitative stability results of section 3 extend earlier work for two-stage models with fixed recourse [18] and for multiperiod two-stage models [10]. Since Theorem 3.2 is stated in terms of the (uniform) semidistances $d_{\mathcal{F}_\rho}$, it allows two types of applications. First, it is possible to utilize metric entropy results and to quantify the asymptotic behavior of statistical approximations to two-stage stochastic programs with random recourse. Second, the analysis of continuity properties of the convex random lsc functions f_0 enables bounding semidistances by appropriate Fortet–Mourier metrics. Such metrics are easier to handle due to their relations to mass transportation problems and their dual representations, particularly for computational purposes (e.g., in scenario reduction algorithms developed in [5, 9]).

The general stability results for model (1.1) in section 2 provide continuity properties of infima and (approximate) solution sets relative to changes of the original probability distribution. They are simple consequences of general perturbation results for optimization problems. Presently, they are stated in terms of the uniform probability semidistance $d_{\mathcal{F}_\rho}$ on the space of probability measures, although the same results would be valid in terms of the corresponding epi-distances $\hat{d}_{\mathcal{I}_\rho}$ or $d_{\mathcal{I}_\rho}$, too. Such epi-distances would allow for richer spaces of probability measures $\mathcal{P}_\mathcal{F}$ and for extended real-valued objective functions $\mathbb{E}^P f_0(x)$ with different effective domains, respectively. But, since a theory for epi-counterparts of uniform distances of Fortet–Mourier type and of uniform large deviation results (see (5.3)) is not yet developed, the achieved generality would appear to be wasted. If, however, these gaps are filled in the future, the framework developed in section 2 forms the basis for extending the present results in sections 3, 4, and 5.

Acknowledgments. The authors wish to thank the guest editor Darinka Dentcheva and an anonymous referee for helpful comments.

REFERENCES

[1] Z. Artstein and R. J.-B. Wets, *Stability results for stochastic programs and sensors, allowing for discontinuous objective functions*, SIAM J. Optim., 4 (1994), pp. 537–550.

[2] Z. Artstein and R. J.-B. Wets, *Consistency of minimizers and the SLLN for stochastic programs*, J. Convex Anal., 2 (1995), pp. 1–17.

[3] H. Attouch and R. J.-B. Wets, *Quantitative stability of variational systems* II. *A framework for nonlinear conditioning*, SIAM J. Optim., 3 (1993), pp. 359–381.

[4] H. Attouch and R. J.-B. Wets, *Quantitative stability of variational systems* III. *ε-approximate solutions*, Math. Programming, 61 (1993), pp. 197–214.

[5] J. Dupačová, N. Gröwe-Kuska, and W. Römisch, *Scenario reduction in stochastic programming: An approach using probability metrics*, Math. Program., 95 (2003), pp. 493–511.

[6] J. Dupačová and R. J.-B. Wets, *Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems*, Ann. Statist., 16 (1988), pp. 1517–1549.

[7] A. Eichhorn and W. Römisch, *Polyhedral risk measures in stochastic programming*, SIAM J. Optim., 16 (2005), pp. 69–95.

[8] R. Fortet and E. Mourier, *Convergence de la répartition empirique vers la répartition théorique*, Ann. Sci. École Norm. Sup., 70 (1953), pp. 266–285.

[9] H. Heitsch and W. Römisch, *A note on scenario reduction for two-stage stochastic programs*, Oper. Res. Lett., to appear.

[10] H. Heitsch, W. Römisch, and C. Strugarek, *Stability of multistage stochastic programs*, SIAM J. Optim., 17 (2006), pp. 511–525.

[11] P. Kall, *On approximations and stability in stochastic programming*, in Parametric Optimization and Related Topics, J. Guddat, H. Th. Jongen, B. Kummer, and F. Nožička, eds., Akademie-Verlag, Berlin, 1987, pp. 387–407.

[12] Y. M. Kaniovski, A. J. King, and R. J.-B. Wets, *Probabilistic bounds (via large deviations) for the solutions of stochastic programming problems*, Ann. Oper. Res., 56 (1995), pp. 189–208.

[13] A. J. King and R. J.-B. Wets, *Epi-consistency of convex stochastic programs*, Stochastics Stochastics Rep., 34 (1991), pp. 83–92.

[14] D. Klatte, *On quantitative stability for non-isolated minima*, Control Cybernet., 23 (1994), pp. 183–200.

[15] G. Pflug, *Stochastic optimization and statistical inference*, in Stochastic Programming, A. Ruszczyński and A. Shapiro, eds., Handbooks Oper. Res. Management Sci. 10, Elsevier, Amsterdam, 2003, pp. 427–482.

[16] G. Pflug, A. Ruszczyński, and R. Schultz, *On the Glivenko–Cantelli problem in stochastic programming: Linear recourse and extensions*, Math. Oper. Res., 23 (1998), pp. 204–220.

[17] S. T. Rachev, *Probability Metrics and the Stability of Stochastic Models*, Wiley, Chichester, UK, 1991.

[18] S. T. Rachev and W. Römisch, *Quantitative stability in stochastic programming: The method of probability metrics*, Math. Oper. Res., 27 (2002), pp. 792–818.

[19] S. M. Robinson and R. J.-B. Wets, *Stability in two-stage stochastic programming*, SIAM J. Control Optim., 25 (1987), pp. 1409–1416.

[20] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[21] R. T. Rockafellar, *Integral functionals, normal integrands, and measurable selections*, in Nonlinear Operators and the Calculus of Variations, J. Gossez and L. Waelbroeck, eds., Lecture Notes in Math. 543, Springer-Verlag, Berlin, 1976, pp. 157–207.

[22] R. T. Rockafellar and S. Uryasev, *Conditional value-at-risk for general loss distributions*, J. Banking & Finance, 26 (2002), pp. 1443–1471.

[23] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, 2nd ed., Springer-Verlag, New York, 2004.

[24] W. Römisch, *Stability of stochastic programming problems*, in Stochastic Programming, A. Ruszczyński and A. Shapiro, eds., Handbooks Oper. Res. Management Sci. 10, Elsevier, Amsterdam, 2003, pp. 483–554.

[25] W. Römisch and R. Schultz, *Stability analysis for stochastic programs*, Ann. Oper. Res., 30 (1991), pp. 241–266.

[26] W. Römisch and R. Schultz, *Lipschitz stability for stochastic programs with complete recourse*, SIAM J. Optim., 6 (1996), pp. 531–547.

[27] W. Römisch and A. Wakolbinger, *Obtaining convergence rates for approximations in stochastic programming*, in Parametric Optimization and Related Topics, J. Guddat, H. Th. Jongen, B. Kummer, and F. Nožička, eds., Akademie Verlag, Berlin, 1987, pp. 327–343.

[28] R. Schultz, *Strong convexity in stochastic programs with complete recourse*, J. Comput. Appl. Math., 56 (1994), pp. 3–22.
[29] A. Shapiro, *Quantitative stability in stochastic programming*, Math. Programming, 67 (1994), pp. 99–108.
[30] A. Shapiro, *Monte Carlo sampling methods*, in Stochastic Programming, A. Ruszczyński and A. Shapiro, eds., Handbooks Oper. Res. Management Sci. 10, Elsevier, Amsterdam, 2003, pp. 353–425.
[31] M. Talagrand, *Sharper bounds for Gaussian and empirical processes*, Ann. Probab., 22 (1994), pp. 28–76.
[32] A. W. van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes*, Springer-Verlag, New York, 1996.
[33] W. Vervaat, *Random Upper Semicontinuous Functions and Extremal Processes*, Report MS-R8801, Center for Wiskunde en Informatica, Amsterdam, 1988.
[34] D. W. Walkup and R. J.-B. Wets, *Lifting projections of convex polyhedra*, Pacific J. Math., 28 (1969), pp. 465–475.
[35] D. Walkup and R. J.-B. Wets, *A Lipschitzian characterization of convex polyhedra*, Proc. Amer. Math. Soc., 23 (1969), pp. 167–173.
[36] R. J.-B. Wets, *Stochastic programs with fixed recourse: The equivalent deterministic program*, SIAM Rev., 16 (1974), pp. 309–339.
[37] R. J.-B. Wets, *On the continuity of the value of a linear program and of related polyhedral-valued multifunctions*, Math. Programming Stud., 24 (1985), pp. 14–29.
[38] R. J.-B. Wets, *Challenges in stochastic programming*, Math. Programming, 75 (1996), pp. 115–135.
[39] R. J.-B. Wets, *Lipschitz continuity of inf-projections*, Comput. Optim. Appl., 25 (2003), pp. 269–282.

SIAM J. OPTIM.
Vol. 18, No. 3, pp. 980–1003

ON A DISTRIBUTED CONTROL PROBLEM ARISING IN DYNAMIC OPTIMIZATION OF A FIXED-SIZE POPULATION*

GUSTAV FEICHTINGER† AND VLADIMIR M. VELIOV‡

Abstract. The practical motivation for this paper is provided by the recruitment problem faced by many organizations of fixed size: to keep the average age young (and thus keep innovation and productivity high) while at the same time keeping levels of recruitment high. A typical example is an academy of sciences. The problem is formalized by an infinite horizon optimal control model for a first order PDE with nonlocal dynamics (a McKendrick-type equation). Based on the nonstandard necessary optimality condition proved in the paper, the following results are established: (i) stationarity of the optimal recruitment density; (ii) strong ergodicity of the optimal solution; (iii) principle of "bipolar" recruitment in the case where the productivity of the organization is measured by the average age of the members. The analysis involves a new type of transversality condition for the costate system, the stability with respect to perturbations of the optimal solution of a noncoercive (bang-bang type) problem, boundedness, and stability of the solution of a specific Volterra integral equation of the second kind.

Key words. optimal control, distributed control, population dynamics, ergodicity, McKendrick equation, infinite horizon problems, stability analysis

AMS subject classifications. 49K20, 49K40, 92D25

DOI. 10.1137/06066148X

1. Introduction.

The dynamics of populations of fixed size plays an important role in demography (e.g., migration to guarantee zero population growth for below-replacement fertility) and in manpower planning; see, e.g., [28, 16, 27, 5, 17, 29]. Optimal investment problems in a fixed-size firm with age-structured physical capital and fixed or variable scrapping age (in the framework of, e.g., [10]) lead to technical issues similar to the ones involved below.

The practical motivation for the fixed-size problem investigated in this paper is the following. It has been observed that in many organizations of fixed size, such as academies of sciences in many countries, the average age of the members has increased in past decades [24]. Therefore, the question of appropriate recruitment strategy arises, i.e., that of ensuring a reasonably low average age of the members of the organization while at the same time providing a reasonably high recruitment rate which is desirable for several reasons.[1] These two objectives turn out to be conflicting [24]. More generally, we consider the average productivity of the organization as one of the objectives to be maximized, the recruitment intensity being the second objective, where we assume that the productivity of the members of the organization

*Received by the editors May 31, 2006; accepted for publication (in revised form) April 12, 2007; published electronically October 4, 2007. This work was supported by the Präsidium of the Austrian Academy of Sciences, and by the Austrian Science Foundation (FWF) under grant P18161-N13.
http://www.siam.org/journals/siopt/18-3/66148.html

†Institute of Mathematical Methods in Economics, Vienna University of Technology, Argentinierstrasse 8, A-1040 Vienna, Austria (gustav.feichtinger@tuwien.ac.at).

‡Institute of Mathematical Methods in Economics, Vienna University of Technology, Argentinierstrasse 8, A-1040 Vienna, Austria (vladimir.veliov@tuwien.ac.at), and Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, 1113 Sofia, Bulgaria.

[1]For an academy of sciences, for example, a too-small number of elections would frustrate the scientists outside the academy and would decrease the stimulative role of the academy. Moreover, the new members bring new ideas, represent new areas, etc.

depends on the age. The following model for the dynamics of the age-structure of the organization involves a nonstandard version of the McKendrick equation [26, 33, 2]:

$$M_t(t,a) + M_a(t,a) = -\mu(t,a)M(t,a) + R(t)u(t,a), \tag{1.1}$$

$$R(t) = M(t,\omega) + \int_0^\omega \mu(t,a)M(t,a)\,\mathrm{d}a, \tag{1.2}$$

with the side conditions

$$M(0,a) = M_0(a), \quad M(t,0) = 0, \tag{1.3}$$

and the state constraint

$$M(t,a) \geq 0.$$

Here $M(t,\cdot)$ is the age-density of the members of the organization at time t; $\mu(t,a)$ is the mortality rate at time t and age a; $R(t)$ is the intensity of recruitment at time t; $u(t,\cdot)$ is the age-density[2] of recruitment at time t; $M_0(\cdot)$ is the initial age-density of members; ω is the (fixed) retirement age of members; and $M_t + M_a$ is the sum of the partial derivatives of M (strictly speaking, this is the derivative of M in the direction $(1,1)$ in the (t,a)-plane).

The dynamics is given by the classical McKendrick equation (1.1), while (1.2) means that the size of the organization is fixed and equals $\bar{M} = \int_0^\omega M_0(a)\,\mathrm{d}a$ (this can be easily seen by integrating (1.1) in a and utilizing the assumption for fixed size[3]).

The following constraints are posed for the recruitment density, $u(t,\cdot)$, which is considered further as a control variable:

$$0 \leq u(t,a) \leq \bar{u}(a), \quad \int_0^\omega u(t,a)\,\mathrm{d}a = 1. \tag{1.4}$$

The upper bound, $\bar{u}(a)$, for the control has a different meaning in different practical situations and is discussed in detail in [14] for the case of an academy of sciences. It will be proved in section 2 that the state constraint $M(t,a) \geq 0$ is automatically satisfied for any admissible control function u; therefore we do not refer to it further.

As mentioned above, we focus our analysis on two objectives that are to be maximized:

- the average productivity of the organization, $\int_0^\omega p(t,a)M(t,a)\,\mathrm{d}a$, and
- the recruitment intensity, $R(t)$.

Here $p(t,a)$ is the productivity of the members of age a. Productivity may have different meanings in different contexts, but for this paper it matters only that it depends on age. The average age of the members is a special case where $p(t,a) = -a/\bar{M}$ (taken with a minus sign since it is to be minimized).

A study of the above multiobjective and state-constrained problem in the spirit of the viability theory for age-structured systems [7] will be presented by the second author elsewhere.

[2]To avoid misunderstanding, we stress that $M(t,\cdot)$ need not be a probabilistic density, while $u(t,\cdot)$ is assumed to be a probabilistic density, in the sense given by the equality in (1.4) below.

[3]The extension for organizations with smoothly changing size is straightforward. The same applies to a time-dependent control constraint in (1.4) below (excepting the results in section 5).

In this paper we employ the Pareto optimization framework, considering the objective function

$$\max \int_0^\infty e^{-rt} \left[\alpha R(t) + \beta \int_0^\omega p(t,a) M(t,a)\, \mathrm{d}a \right] \mathrm{d}t, \tag{1.5}$$

where $r > 0$ is a time-preference rate, and $\alpha > 0$ and $\beta > 0$ are weights attributed to the two objectives.

Although the above problem is a rather specific one, it is of substantial interest and reflects the intensive policy-oriented discussions that have taken place in many academies of sciences during the past few years [24, 15]. What is more important for the present paper is that the problem provides a number of challenges: (i) it is nonlinear (although bilinear); (ii) the time horizon is infinite, which creates substantial difficulties in obtaining appropriate transversality conditions for the costate system which are strong enough to facilitate the stability analysis needed for the main results[4]; (iii) due to (1.2) the dynamics in (1.1) is nonlocal; and (iv) the optimal solution is of bang-bang type, which is known to create substantial difficulties in the stability analysis of the optimal control also for ODEs (cf., e.g., [23, 19]).

The main results are (i) time-invariance of the optimal recruitment distribution in the case of time-invariant data (section 5); (ii) strong ergodicity of the optimal solution (section 6); and (iii) characterization of the optimal control in a case of particular practical interest by the *principle of bipolar recruitment* (section 7).

All these results are based on the nonstandard optimality condition of Pontryagin type obtained in section 3, with a rather strong transversality condition, which is crucial for the main result (compare with the transversality conditions in the state-of-the-art paper [6]). The proof of the transversality condition is given in the appendix. Some basic properties of the problem under consideration are presented in section 2, and a basic auxiliary result is proved in section 4.

2. Basic properties. In this section we consider a somewhat more general objective function which arises in other models of organizations with fixed size:

$$\max \int_0^\infty e^{-rt} \left[\alpha R(t) + \beta \int_0^\omega A(t,a,M(t,a),R(t)u(t,a))\, \mathrm{d}a \right] \mathrm{d}t, \tag{2.1}$$

subject to (1.1)–(1.4) (the sign "∞" means everywhere "$+\infty$"). Assuming A is dependent on the size of the inflow $w = Ru$ is reasonable due to possible adjustment costs.

Denote for brevity $\Omega = [0,\omega]$, $D = [0,\infty) \times [0,\omega]$.

Standing assumptions. $\mu : D \mapsto [0,\bar{\mu}]$ is measurable; $\bar{u} : \Omega \mapsto [0,\bar{v}]$ is measurable, $\int_\Omega \bar{u}(a)\, \mathrm{d}a > 1$; $M_0 : \Omega \mapsto [0,\infty)$ is measurable and bounded; $A : D \times \mathbf{R} \times \mathbf{R}$ is bounded (locally in (M,w)), measurable in t,a, concave in (M,w), and differentiable in M and w; and the derivatives A_M and A_w are bounded and Lipschitz continuous in (M,w) uniformly in (t,a). Here $\bar{\mu} \geq 0$ and $\bar{v} > 0$ are constants.

By definition, $M \in L_\infty^{\mathrm{loc}}(D)$ is a solution of (1.1) (for given measurable functions R and u) if it can be represented by a function which is absolutely continuous on

[4] Out of many papers that consider optimal control of McKendrick equations, only [25] studies the asymptotic behavior of the adjoint variable for a special system (not including the problem considered here), but the transversality condition obtained there is based on an implicit assumption, the verification of which is, actually, the main trouble. Our approach is substantially different.

almost every (a.e.) characteristic line $t - a = \text{const}$ and satisfies (1.1) almost everywhere on a.e. characteristic line, where the symbol $M_t + M_a$ is interpreted as the directional derivative in direction $(1, 1)$. Notice that the trace of a solution, M, on every straight line which is transversal to the characteristic lines is well defined; in particular, $M(\cdot, \omega)$, $M(0, \cdot)$, $M(t, 0)$, etc. are well-defined elements of L_∞^{loc}. Consequently, the side conditions (1.3) are understood as equalities in L_∞^{loc} (more details are given in [18]; see also [33, 2]). Equation (1.2) has the meaning of an equality between L_∞-functions.

LEMMA 2.1. *For every measurable function u satisfying* (1.4)*, the system* (1.1)–(1.3) *has a unique solution in D; the solution is essentially bounded (uniformly in u) and satisfies $M(t, a) \geq 0$.*

Proof. For a given function $R \in L_\infty^{\mathrm{loc}}(0, \infty)$, the solution of (1.1), (1.3) has the following explicit representation, obtained by application of the Cauchy formula on the characteristic lines $t - a = \text{const}$:

$$M(t,a) = \psi(0,a,t)\chi_\Omega(a-t)M_0(a-t) + \int_0^t \psi(\tau,a,t)\chi_\Omega(a-t+\tau)R(\tau)u(\tau,a-t+\tau)\,\mathrm{d}\tau, \tag{2.2}$$

where χ_Ω is the characteristic function of the set Ω, and

$$\psi(\tau,a,t) = e^{-\int_\tau^t \mu(\theta,a-t+\theta)\,\mathrm{d}\theta}. \tag{2.3}$$

Plugging this into (1.2) and changing the order of integration we obtain the following equation for R:

$$R(t) = f(t) + \int_0^t K(t,\tau)R(\tau)\,\mathrm{d}\tau, \tag{2.4}$$

where

$$\begin{aligned} f(t) &= \psi(0,\omega,t)\chi_\Omega(\omega-t)M_0(\omega-t) + \int_\Omega \mu(t,a)\psi(0,a,t)\chi_\Omega(a-t)M_0(a-t)\,\mathrm{d}a, \\ K(t,\tau) &= \psi(\tau,\omega,t)\chi_\Omega(\omega-t+\tau)u(\tau,\omega-t+\tau) \\ &\quad + \int_\Omega \mu(t,a)\psi(\tau,a,t)\chi_\Omega(a-t+\tau)u(\tau,a-t+\tau)\,\mathrm{d}a. \end{aligned}$$

Since this is a Volterra integral equation of the second kind, it has a solution in $L_\infty^{\mathrm{loc}}(0, \infty)$ (see, e.g., [20, Theorem 4.2, Chapter 9]). It remains to prove that $R \in L_\infty(0, \infty)$, which is the main message of the lemma.

As already mentioned in the introduction, (1.2) is equivalent to

$$\int_0^\omega M(t,a)\,\mathrm{d}a = \bar{M}, \quad \text{where } \bar{M} = \int_0^\omega M_0(a)\,\mathrm{d}a.$$

Due to the conditions (1.4) for u, there exists $\theta > 0$ such that $\int_0^{\omega-\theta} u(t,s)\,\mathrm{d}s \geq 1/2$ for every t. Then, using (2.2), we have

$$\bar{M} \geq \int_0^\omega \int_0^t \psi(\tau,a,t)\chi_\Omega(a-t+\tau)R(\tau)u(\tau,a-t+\tau)\,\mathrm{d}\tau\,\mathrm{d}a.$$

If we extend R as $R(t) = 0$ for $t < 0$, we have

$$\bar{M} \geq \int_{t-\omega}^{t} \int_{t-\tau}^{\omega} \psi(\tau, a, t) R(\tau) u(\tau, a - t + \tau) \, \mathrm{d}a \, \mathrm{d}\tau \geq c \int_{t-\omega}^{t} \int_{0}^{\omega-t+\tau} R(\tau) u(\tau, s) \, \mathrm{d}s \, \mathrm{d}\tau,$$

where c is a lower bound for $\psi(\tau, a, t)$, which is positive since $t - \tau \leq \omega$. Hence,

$$\bar{M} \geq c \int_{t-\theta}^{t} \int_{0}^{\omega-t+\tau} R(\tau) u(\tau, s) \, \mathrm{d}s \, \mathrm{d}\tau \geq \frac{c}{2} \int_{t-\theta}^{t} R(\tau) \, \mathrm{d}\tau.$$

Since this inequality holds for every $t > 0$, we have

$$\int_{t-\omega}^{t} R(\tau) \, \mathrm{d}\tau \leq \int_{t-\theta}^{t} + \int_{t-2\theta}^{t-\theta} + \cdots + \int_{t-k\theta}^{t-(k-1)\theta} \leq \frac{2k}{c} \bar{M},$$

where k is the smallest natural number satisfying $k\theta \geq \omega$. Then formula (2.2), in which the integration is, in fact, carried out on the interval $[t - a, t] \subset [t - \omega, t]$ (due to the presence of the characteristic function of Ω), implies that

$$M(t, a) \leq \|M_0\|_{L_\infty(\Omega)} + \frac{2k\bar{u}}{c} \bar{M} \quad \text{almost everywhere.}$$

The essential boundedness of R follows from (1.2).

To prove that $M(t, a) \geq 0$ we note that $K(t, a) \geq 0$ and $f(t) \geq 0$; hence $R(t) \geq 0$ [20, Proposition 8.1, Chapter 9]. Then (2.2) implies $M(t, a) \geq 0$. □

PROPOSITION 2.1. *Problem* (1.1)–(1.4), (2.1) *has a solution. In particular, problem* (1.1)–(1.5) *has a solution.*

The proof is given in the appendix. Although the idea of the proof is applicable only to a rather restricted class of problems, it could be useful beyond the specific problem (1.1)–(1.4), (2.1).

Uniqueness of the optimal solution is proved under additional conditions in Theorem 5.1. In general, uniqueness is not granted, as Example 1 in section 5 shows. Here we mention only that the function "admissible control" $\longrightarrow$ "objective value" is not necessarily concave with respect to the control[5]; therefore, the uniqueness requires some additional assumptions such as the one formulated in section 5.

3. Optimality conditions. Problem (1.1)–(1.4), (2.1) is on the infinite horizon and involves the "advanced" term $M(t, \omega)$ in the right-hand side of the differential equation. For each of these two reasons the maximum principle obtained in [11, 18], and the other known optimality conditions for McKendrick-type control systems (see the references in [18] and footnote 4) are not applicable.

[5]The claim that the function "admissible control u" $\longrightarrow$ "objective value $J(u)$" is not concave is nonobvious. We have a (generic) counterexample with $\mu(a) = 0$, $A = -aM$, but the argument behind it is too long, and we skip it due to the page limitation. The main idea is first to consider the steady-state version of the problem, where (1.1)–(1.3) (with $M_t = 0$) are analytically solvable. This allows us to represent $J(u)$ in the form

$$J(u) = \frac{c + d\mu_2(u)}{\omega - \mu_1(u)}, \quad c, d > 0,$$

where $\mu_1(u)$ and $\mu_2(u)$ are the first and the second integral moments of u. It is easy to see that the above function is strictly convex on the segment between two controls u_1 and u_2 for which $\mu_1(u_1) = \mu_2(u_1)$ and $\mu_1(u_2) = \mu_2(u_2)$. After that we consider the intertemporal problem (1.1)–(1.4), (2.1) and argue that the nonconcavity is preserved there if the discount r is sufficiently small.

To obtain a necessary optimality condition we introduce the following *adjoint system* for given reference L_∞-functions M and u:

$$\xi_t(t,a)+\xi_a(t,a)=(r+\mu(t,a))\xi(t,a)-\mu(t,a)\eta(t) - \beta A_M(t,a,M(t,a),R(t)u(t,a)), \tag{3.1}$$

$$\eta(t)=\alpha+\int_0^\omega \xi(t,a)u(t,a)\,\mathrm{d}a+\int_0^\omega A_w(t,a,M(t,a),R(t)u(t,a))u(t,a)\,\mathrm{d}a, \tag{3.2}$$

with the boundary condition

$$\xi(t,\omega)=\eta(t). \tag{3.3}$$

The meaning of a solution is the same as for the primal system (1.1)–(1.3). However, the issue of existence and uniqueness here is much more complicated. Actually the above system has infinitely many solutions in $L_\infty^{\mathrm{loc}}(D)$, while it has exactly one solution belonging to $L_\infty(D)$. The proof of the next lemma is given in the appendix.

LEMMA 3.1. *Let u be a control function satisfying* (1.4). *Then system* (3.1)–(3.3) *has a unique solution (ξ,η) in the space $L_\infty(D)\times L_\infty(0,\infty)$.*

Remark 3.1. In general, certain transversality conditions are needed to ensure uniqueness of the solution of the adjoint system for infinite horizon optimal control problems. The usual form of the transversality condition for a discounted objective integrand (see [6]) adapted to our problem would be $\lim_{t\to\infty} e^{-rt}\|\xi(t,\cdot)\|_{L_\infty(\Omega)}=0$. Notice that the condition $\xi\in L_\infty(D)$ represents a stronger transversality condition which obviously implies the above one. This strong transversality condition plays a key role in the subsequent analysis.[6]

THEOREM 3.1. *Let (u,M,R) be an optimal solution, and let ξ be the unique solution of the adjoint system* (3.1)–(3.3) *in $L_\infty(D)$. Then for a.e. t the optimal control, $u(t,\cdot)$, maximizes the integral*

$$\int_0^\omega [\xi(t,a)R(t)v(a)+\beta A(t,a,M(t,a),R(t)v(a))]\,\mathrm{d}a$$

on the set of functions $v(\cdot)$ satisfying

$$0\le v(a)\le \bar u(a),\quad \int_0^\omega v(a)\,\mathrm{d}a=1. \tag{3.4}$$

The main difficulty is encapsulated in Lemma 3.1; therefore, below we only sketch the rest of the proof, which is a matter of technical manipulations with an appropriate needle variation.

Proof. Let us fix an arbitrary $s>0$ which is a Lebesgue point of the functions

$$t\longrightarrow\int_\Omega \xi(t,a)R(t)u(t,a)\,\mathrm{d}a \quad\text{and}\quad t\longrightarrow\int_\Omega A(t,a,M(t,a),u(t,a))\,\mathrm{d}a,$$

and also of every function of the form

$$t\longrightarrow\int_\Omega \xi(t,a)R(t)\tilde u(a)\,\mathrm{d}a \quad\text{and}\quad t\longrightarrow\int_\Omega A(t,a,M(t,a),\tilde u(a))\,\mathrm{d}a,$$

[6] One of the referees suggested that boundedness of the adjoint variable might be possible to prove also using the Lipschitz continuity of the value function of the problem and employing the Barron–Jensen [9, 8] approach for proving the maximum principle.

where $\tilde{u} \in L_\infty(\Omega)$ satisfies (3.4). Elementary measure-theoretic considerations, together with the separability of $L_1(\Omega)$ and the boundedness of R, M, and ξ, yield that a.e. s is a Lebesgue point of all these functions. The main point is to observe (we skip the easy proof of this) that if s is a Lebesgue point of the last two groups of functions for any $u = u_j$ belonging to a countable dense set in $L_1(\Omega)$, then it is a Lebesgue point for these functions for every $u \in L_1(\Omega)$ also.

Denote $T = s + \omega$. Let $h \in (0, s) \cap \Omega$ be a "small" parameter, and denote $S_h := [s - h, s + h] \times \Omega$. Let v be an arbitrary measurable function satisfying (3.4). Consider the variation

$$\Delta u(t,a) = \begin{cases} 0 & \text{for } (t,a) \notin S_h, \\ v(a) - u(t,a) & \text{for } (t,a) \in S_h. \end{cases}$$

Obviously $u+\Delta u$ satisfies (3.4) and therefore is an admissible control. For the solution $(M + \Delta M, R + \Delta R)$ of (1.1)–(1.3) corresponding to $u + \Delta u$ it holds that

$$\mathcal{D}\Delta M = -\mu\Delta M + \Delta R(t)(u+\Delta u) + R(t)\Delta u, \quad \Delta M(0,a) = 0, \quad \Delta M(t,0) = 0, \tag{3.5}$$

$$\Delta R(t) = \Delta M(t,\omega) + \int_0^\omega \mu(t,a)\Delta M(t,a)\,\mathrm{d}a, \tag{3.6}$$

where $\mathcal{D}$ is the derivative in the direction $(1,1)$, and we skip the arguments (t,a). Applying formula (2.2) to (3.5) and then substituting ΔM in (3.6) we obtain (as in the proof of Lemma 2.1) a Volterra equation of the second kind for ΔR:

$$\Delta R(t) = f(t) + \int_0^t K(t,\tau)\Delta R(\tau)\,\mathrm{d}\tau,$$

where $K(t,\tau) \le C(s)$ (here and below $C(s), C_1(s), \dots$ denote numbers that may depend on s but not on h) for $0 \le \tau \le t \le T$, and

$$\begin{aligned} f(t) = &\int_0^t \psi(\tau,\omega,t)\chi_\Omega(\omega - t + \tau)R(\tau)\Delta u(\tau,\omega - t + \tau)\,\mathrm{d}\tau \\ &+ \int_0^t \int_0^\omega \mu(t,a)\psi(\tau,a,t)\chi_\Omega(a - t + \tau)R(\tau)\Delta u(\tau,a - t + \tau)\,\mathrm{d}\tau\,\mathrm{d}a. \end{aligned}$$

From here one can estimate

$$|f(t)| \le C_1 h.$$

Since the above Volterra equation has a locally bounded resolvent kernel (see, e.g., [20, Corollary 4.3, Chapter 9]), $H(t,\tau)$, we may express

$$\Delta R(t) = f(t) + \int_0^t H(t,\tau)f(\tau)\,\mathrm{d}\tau.$$

Hence $|\Delta R(t)| \le C_2(s)h$. Then we have

$$\int_D |\Delta R(t)\Delta u(t,a)|\,\mathrm{d}(t,a) \le \int_{S_h} |\Delta R(t)\Delta u(t,a)|\,\mathrm{d}(t,a) = O_s(h^2),$$

where $O_s(h^2) \le C_3(s)h^2$.

We multiply (3.5) by $e^{-rt}\xi(t,a)$ and integrate on D:

$$\int_D e^{-rt}\xi(t,a)\mathcal{D}\Delta M(t,a)\,\mathrm{d}(t,a)$$
$$= \int_D e^{-rt}\xi(t,a)[-\mu(t,a)\Delta M(t,a) + \Delta R(t)u(t,a) + R(t)\Delta u(t,a)]\,\mathrm{d}(t,a) + O_s(h^2).$$

Due to the relations $\Delta M(t,0) = 0$, $\Delta M(0,a) = 0$, and $\xi \in L_\infty(D)$, simple calculations that we skip transform the left-hand side to

$$\int_0^\infty \xi(t,\omega)\Delta M(t,\omega)\,\mathrm{d}t - \int_D e^{-rt}\left[-r\xi(t,a) + \mathcal{D}\xi(t,a)\right]\Delta M(t,a)\,\mathrm{d}(t,a). \tag{3.7}$$

Due to (3.6), the right-hand side becomes

$$\int_D e^{-rt}\left[-\mu(t,a)\xi(t,a) + \mu(t,a)\int_\Omega \xi(t,b)u(t,b)\,\mathrm{d}b\right]\Delta M(t,a)\,\mathrm{d}(t,a) \tag{3.8}$$

$$+ \int_D e^{-rt}\xi(t,a)\Delta M(t,\omega)u(t,a)\,\mathrm{d}(t,a) + \int_D e^{-rt}\xi(t,a)R(t)\Delta u(t,a)\,\mathrm{d}(t,a) + O_s(h^2).$$

Since u is optimal, for the difference of the objective values, $J(u+\Delta u) - J(u) \le 0$, we have

$$0 \ge \int_0^\infty e^{-rt}\alpha\Delta M(t,\omega)\,\mathrm{d}t$$
$$+ \int_0^\infty e^{-rt}\int_0^\omega \left[\alpha\mu(t,a)\Delta M(t,a) + \beta A_M(t,a)\Delta M(t,a) + \beta A_w(t,a)\Delta R(t)u(t,a)\right.$$
$$\left. +\beta\left(A(t,a,M(t,a),R(t)(u(t,a)+\Delta u(t,a))) - A(t,a)\right)\right]\,\mathrm{d}a\,\mathrm{d}t + O_s(h^2),$$

where here and below the argument (t,a) of the functions A, A_w, and A_w is a substitution for $(t,a,M(t,a),R(t)u(t,a))$. We subtract from the right-hand side of the above inequality the expression in (3.7) and add the expression (3.8) which has the same value. Rearranging the terms and taking into account that ξ is the solution of the adjoint system (3.1)–(3.3), we obtain

$$O_s(h^2) \ge \int_D e^{-rt}[\xi(t,a)R(t)\Delta u(t,a)$$
$$+\beta\left(A(t,a,M(t,a),R(t)(u(t,a)+\Delta u(t,a))) - A(t,a)\right)]\,\mathrm{d}(t,a).$$

This is equivalent to

$$\frac{1}{2h}\int_{s-h}^{s+h} e^{-rt}\int_0^\omega \left[\xi(t,a)R(t)u(t,a) + \beta A(t,a,M(t,a),R(t)u(t,a))\right]\,\mathrm{d}a\,\mathrm{d}t$$
$$\ge \frac{1}{2h}\int_{s-h}^{s+h} e^{-rt}\int_0^\omega \left[\xi(t,a)R(t)v(a) + \beta A(t,a,M(t,a),R(t)v(a))\right]\,\mathrm{d}a\,\mathrm{d}t - O_s(h).$$

Then the claim of the theorem follows by taking the limit with $h \to 0$ and using the definition of a Lebesgue point. □

4. An auxiliary result. From now on we consider only the original problem (1.1)–(1.5).

Since for the objective (1.5) the function $A = p(t,a)M$ is independent of u, whenever $R(t) > 0$ we may equivalently rewrite the maximization condition in Theorem 3.1 as

$$\int_0^\omega \xi(t,a)u(t,a)\,\mathrm{d}a = \max_{v:(3.4)} \int_0^\omega \xi(t,a)v(a)\,\mathrm{d}a, \tag{4.1}$$

where (as indicated) the maximization is carried out on the set of functions $v \in L_\infty(\Omega)$ satisfying (3.4). If $R(t) = 0$ the value of u is of no meaning. For convenience we define in this case $u(t,\cdot)$ as an arbitrary maximizer in (4.1), so that with this convention (4.1) holds for every optimal control for a.e. t.[7]

Denote

$$\sigma(\lambda) = \max_v \int_0^\omega \lambda(a)v(a)\,\mathrm{d}a, \quad \lambda \in L_\infty(\Omega).$$

Then the adjoint system (3.1) can be rewritten in the following closed (feedback) form:

$$\xi_t + \xi_a = (r + \mu(t,a))\xi - \mu(t,a)(\alpha + \sigma(\xi(t,\cdot))) - \beta p(t,a), \tag{4.2}$$

$$\xi(t,\omega) = \alpha + \sigma(\xi(t,\cdot)). \tag{4.3}$$

The existence of a solution to (1.1)–(1.5) and the necessity of the maximum principle imply that the above functional-differential system has at least one solution in $L_\infty(D)$. It will be useful to study the stability of the solution with respect to perturbations in the right-hand side.

LEMMA 4.1. *Let $\xi \in L_\infty(D)$ be a solution of* (4.2), (4.3). *Let $\xi^\delta \in L_\infty(D)$ be a solution of* (4.2), (4.3) *(if such exists) with a perturbation $\delta \in L_\infty(D)$ added to the right-hand side of* (4.2). *Then for every $T > 0$*

$$\|\xi^\delta - \xi\|_{L_\infty([T,\infty)\times\Omega)} \le \frac{\|\delta\|_{L_\infty([T,\infty)\times\Omega)}}{1-q},$$

where $q \in (0,1)$ is a number depending only on ω, r, and μ.

Proof. First we shall prove that σ is Lipschitz continuous with a Lipschitz constant equal to one. Indeed, for λ_1, $\lambda_2 \in L_\infty(\Omega)$ we have (having in mind (3.4))

$$\sigma(\lambda_1) = \max_v \int_\Omega \lambda_1(a)v(a)\,\mathrm{d}a = \max_v \left[\int_\Omega \lambda_2(a)v(a)\,\mathrm{d}a + \int_\Omega (\lambda_1(a) - \lambda_2(a))v(a)\,\mathrm{d}a\right]$$
$$\le \max_v \int_\Omega \lambda_2(a)v(a)\,\mathrm{d}a + \max_v \int_\Omega |\lambda_1(a) - \lambda_2(a)|v(a)\,\mathrm{d}a = \sigma(\lambda_2) + \|\lambda_1 - \lambda_2\|_{L_\infty(\Omega)}.$$

If a solution ξ^δ exists, we consider $\xi^\delta(t,\omega)$ as known, and solve (4.2) along the characteristic lines, which gives for a.e. (t,a)

$$\xi^\delta(t,a) = \varphi(t,a,t+\omega-a)(\alpha + \sigma(\xi^\delta(t+\omega-a,\cdot)))$$

[7] This convention allows us to claim uniqueness of the optimal control later, without excluding the possibility that $R(t) = 0$ for some t. Without this convention the uniqueness of the optimal control proved later should be understood in the sense that for every optimal $u(t,a)$ the corresponding function $R(t)$ is the same, and the values $u(t,a)$ are also the same if $R(t)$ is strictly positive.

$$+\int_t^{t+\omega-a}\varphi(t,a,\tau)\big[\mu(\tau,a-t+\tau)(\alpha+\sigma(\xi^\delta(\tau,\cdot)))$$
$$+\beta p(\tau,a-t+\tau)-\delta(\tau,a-t+\tau)\big]\,\mathrm{d}\tau,$$

where φ is defined in (A.5) in the appendix.

Using the above formula also for $\delta=0$ and taking the difference, we obtain for a.e. $t\ge T$ and $a\in\Omega$ that

$$|\xi^\delta(t,a)-\xi(t,a)|\le\varphi(t,a,t+\omega-a)|\sigma(\xi^\delta(t+\omega-a,\cdot))-\sigma(\xi(t+\omega-a,\cdot))|$$
$$+\int_t^{t+\omega-a}\varphi(t,a,\tau)\left[\mu(\tau,a-t+\tau)|\sigma(\xi^\delta(\tau,\cdot))-\sigma(\xi^\delta(\tau,\cdot))|+|\delta(\tau,a-t+\tau)|\right]\,\mathrm{d}\tau$$
$$\le\left[\varphi(t,a,t+\omega-a)+\int_t^{t+\omega-a}\varphi(t,a,\tau)\mu(\tau,a-t+\tau)\,\mathrm{d}\tau\right]\|\xi^\delta-\xi\|_{L_\infty([t,t+\omega]\times\Omega)}$$
$$+\int_t^{t+\omega-a}|\delta(\tau,a-t+\tau)|\,\mathrm{d}\tau\le q\|\xi^\delta-\xi\|_{L_\infty([t,t+\omega]\times\Omega)}+\|\delta\|_{L_\infty([T,\infty)\times\Omega)},$$

where

$$q=\varphi(t,a,t+\omega-a)+\int_t^{t+\omega-a}\varphi(t,a,\tau)\mu(\tau,a-t+\tau)\,\mathrm{d}\tau.$$

Since from (A.5) we have

$$\varphi_\tau(t,a,\tau)=-\varphi(t,a,\tau)(r+\mu(\tau,a-t+\tau)),$$

it is an elementary exercise to prove that

$$q\le 1-\frac{r}{r+\bar\mu}\left(1-e^{-\omega(r+\bar\mu)}\right)<1.$$

Denote $\varepsilon=\|\delta\|_{L_\infty([T,\infty)\times\Omega)}$. Since the above (t,a) are arbitrary (with $t>T$), we estimate successively

$$|\xi^\delta(t,a)-\xi(t,a)|\le\varepsilon+q\|\xi^\delta-\xi\|_{L_\infty([t,t+\omega]\times\Omega)}\le\varepsilon+q(\varepsilon+q\|\xi^\delta-\xi\|_{L_\infty([t,t+2\omega]\times\Omega)})$$
$$\le\cdots\le\varepsilon+q(\varepsilon+q(\varepsilon+(\ldots)))\le\frac{\varepsilon}{1-q}.$$

Here we have used that $\xi^\delta-\xi\in L_\infty(D)$, so that $q^k\|\xi^\delta-\xi\|_{L_\infty([t,t+k\omega]\times\Omega)}\to 0$. The above estimation implies the lemma. □

COROLLARY 4.1. *Equations* (4.2), (4.3) *have a unique solution in* $L_\infty(D)$.

Proof. It is enough to apply the above lemma with $\delta=0$. □

5. Time-invariance and stability of the optimal control for stationary data. In this section we prove that in the case of stationary (independent of t) data μ and p, the optimal control of problem (1.1)–(1.5) is also stationary. (Of course, the recruitment intensity, $R(t)$, and the state density, $M(t,\cdot)$, are time-dependent in general.) This (at first glance) surprising property is a consequence of the linearity of the system with respect to the state, M, and of the stability with respect to perturbations of the closed-loop adjoint system (4.2), (4.3), established in Lemma 4.1. We also prove uniqueness of the optimal control and stability with respect to data perturbations, which plays a crucial role for the ergodicity theorem in the next section. Notice that the problem we consider has a bang-bang type solution; therefore,

the issue of stability of the optimal control is complicated (cf., e.g., the recent papers [19, 32]). Below we introduce an additional condition that implies the stability and, in fact, also implies sufficiency of the maximum principle, an issue which is still under investigation, also for bang-bang type ODEs (see, e.g., [31, 1, 23]).

Regularity assumption. For all real numbers $d > 0$ and e it holds that

$$\operatorname{meas}\{a \in \Omega : \ \mu(a) + dp(a) = e\} = 0.$$

Important cases where the regularity assumption is fulfilled are discussed in section 7.

THEOREM 5.1. *Let the regularity assumption hold. Let $\mu(t,a) = \mu(a)$ and $p(t,a) = p(a)$ be time-invariant. Then the optimal control for problem* (1.1)–(1.5) *is unique (see footnote* 7*) and time-invariant (that is, $u(t,a) = u(a)$) and satisfies the following two conditions:*

$$\text{(5.1)} \qquad \text{(i)} \qquad \int_\Omega \lambda(a)u(a)\,\mathrm{d}a = \max_v \int_\Omega \lambda(a)v(a)\,\mathrm{d}a,$$

subject to (3.4)*, and*

$$\text{(5.2)} \qquad \text{(ii)} \qquad \int_\Omega \lambda(a)u(a)\,\mathrm{d}a = -\alpha,$$

where λ is the solution of the equation

$$\text{(5.3)} \qquad \dot\lambda = (r + \mu(a))\lambda - \beta p(a) + \nu, \quad \lambda(\omega) = 0,$$

and ν is a "free" parameter.

In fact, the free parameter, ν, in the above formulation should be determined in such a way that the solution of (5.1), (3.4) for this value of ν also satisfies the equality (5.2) (with λ solving (5.3)). It is worth mentioning that the theorem implies that the optimal control is independent of the initial condition M_0. As a consequence of the linearity of (1.1), (1.2), where the optimal u is plugged, the discounted value function of the problem is linear. This fact does not seem to be obvious a priori.

Proof. For stationary data μ and p the adjoint system (4.2), (4.3) is also stationary. This alone does not imply that any solution is stationary, but below we prove first that a stationary solution $\xi(t,a) = \xi(a)$ exists.

Consider the equation

$$\text{(5.4)} \quad \xi_a(a) = (r + \mu(a))\xi(a) - \mu(a)(\alpha + \sigma(\xi(\cdot))) - \beta p(a), \quad \xi(\omega) = \alpha + \sigma(\xi(\cdot)).$$

Denote $\eta = \alpha + \sigma(\xi(\cdot))$. For a fixed (given) η the unique solution of the above equation can be explicitly written by the Cauchy formula. Replacing the expression for ξ in the implicit end-point condition $\xi(\omega) = \alpha + \sigma(\xi(\cdot))$, we obtain one linear equation for η: $b\eta = c$, where the coefficient b has the form

$$b = 1 - \int_\Omega \left[\varphi(a,a,\omega)u(a) - \int_a^\omega \varphi(a,a,\tau)\mu(\tau)u(\tau)\,\mathrm{d}\tau\right] \mathrm{d}a$$

(see (A.5) for the definition of φ). Using the specific form of φ and the inequality $r > 0$ one can easily prove that the above expression is strictly positive. This proves existence of a solution, ξ, of (5.4).

Obviously the extension of $\xi(a)$ on D as a time-invariant function, $\xi(t,a)=\xi(a)$, satisfies the adjoint system (4.2), (4.3), and according to Corollary 4.1, it is its unique solution. Then the maximum principle (Theorem 3.1) claims that for every optimal control and for a.e. fixed t, the function $u(a)=u(t,a)$ satisfies for a.e. t the conditions

$$\int_\Omega \xi(a)u(a)\,\mathrm{d}a = \max_{v:(3.4)} \int_\Omega \xi(a)v(a)\,\mathrm{d}a = \sigma(\xi). \tag{5.5}$$

For the time-invariance of $u(t,a)$ it remains to prove that (5.5) has a unique solution, $u(a)$ (that is, $u(t,a)$ must be the same for all t).

Introducing the function $\lambda(a)=\xi(a)-\eta$ (where $\eta=\alpha+\sigma(\xi(\cdot))$) we easily see that (5.3) (with $\nu=r\eta$) is satisfied, and equalities (5.2) and (5.1) follow from (5.5). Corollary 5.2 formulated below claims the uniqueness of the solution of (5.5) and completes the proof. □

COROLLARY 5.1. *Let the measurable and bounded functions λ and u on Ω satisfy conditions* (i) *and* (ii) *in Theorem* 5.1. *Then there is $l<0$ such that the optimal control $u(a)$ has the following structure:*

$$u(a)=0 \ \text{ for } \ a\in\Omega_-(l), \quad u(a)=\bar u(a) \ \text{ for } \ a\in\Omega_+(l), \tag{5.6}$$

where $\Omega_-(l)=\{a\in\Omega:\ \lambda(a)<l\}$, $\Omega_+(l)=\{a\in\Omega:\ \lambda(a)>l\}$.

Proof. We may apply the Kuhn–Tucker-type result in Theorem 4 [21, 22, Chapter 1] (in the space $L_1(\Omega)$). The inequality for $\bar u$ in the standing assumptions imply that the Lagrange multiplier to the objective function can be taken equal to one. Then the theorem claims that there exists $l\in\mathbf{R}$ (a Lagrange multiplier for the equality constraint in (3.4)) such that u solves the problem

$$\max_v \int_\Omega [\lambda(a)v(a)-lv(a)]\,\mathrm{d}a$$

subject to $0\le v(a)\le\bar u(a)$. Then (5.6) is obvious. Moreover, if $l>0$ then $u(a)$ may happen only if $\lambda(a)\ge 0$ and (5.2) cannot be fulfilled. □

Clearly, if λ is constant on some subset of Ω, then it may happen that $u(a)$ is not uniquely determined for such a and furthermore may have an arbitrary value in $[0,\bar u(a)]$. This will be the case in Example 1 given at the end of the section.

LEMMA 5.1. *Let the regularity assumption be fulfilled. Let λ be a solution of* (5.3) *(with some ν) and u be a solution of the problem* (5.1), (3.4). *Assume that* (5.2) *is fulfilled. Then for every $\varepsilon>0$ there exists $\delta>0$ such that for every $\tilde\lambda\in L_\infty(\Omega)$ for which $\|\tilde\lambda-\lambda\|_{L_\infty(\Omega)}<\delta$ and for every corresponding solution, $\tilde u$, of* (5.1), (3.4) *(with $\tilde\lambda$ substituted for λ) it holds that*

$$\mathrm{meas}\{a\in\Omega:\ \tilde u(a)\ne u(a)\}<\varepsilon.$$

Proof. Assume that the claim is false. Then there exist $\varepsilon>0$ and a sequence $\lambda_k\in L_\infty(\Omega)$ such that $\|\lambda_k-\lambda\|_{L_\infty(\Omega)}<1/k$, and there are corresponding solutions u_k of (5.1), (3.4) such that

$$\mathrm{meas}\{a\in\Omega:\ u_k(a)\ne u(a)\}\ge\varepsilon.$$

We apply Corollary 5.1 to u. The regularity assumption implies that the set $\Omega_0(l)=\Omega\setminus(\Omega_-(l)\cup\Omega_+(l))$ has measure zero. Indeed, in the opposite case there would exist a subset of $\Omega_0(l)$ on which $\lambda'(a)$ exists and equals zero. That is, $(r+\mu(a))l-\beta p(a)+\nu=0$

on a set of positive measure, which contradicts the regularity assumption since $l < 0$; hence $d = -\beta/l > 0$.

Applying Corollary 5.1 to u_k and λ_h, we define the sets $\Omega^k_-(l_k)$ and $\Omega^k_+(l_k)$ determining the structure of u_k. Then we have

$$\{a \in \Omega :\ u_k(a) \neq u(a)\} \subset \Omega_0(l) \cup (\Omega_-(l) \setminus \Omega^k_-(l_k)) \cup (\Omega_+(l) \setminus \Omega^k_+(l_k)). \tag{5.7}$$

If $\Omega_-(l) \setminus \Omega^k_-(l_k)$ has a positive measure, than on this set $l > \lambda(a) \geq \lambda_k(a) - 1/k \geq l_k - 1/k$. Similarly, if $\Omega_+(l) \setminus \Omega^k_+(l_k)$ is of positive measure, then $l < l_k + 1/k$.

Since the set in the right-hand side of (5.7) has a measure at least ε, and $\Omega_0(l)$ is of measure zero, at least one of the other two sets in (5.7) is of positive measure. However, since u and u_k satisfy the equality in (3.4), a standard measure-theoretic exercise shows that if one of the last two sets in (5.7) is of positive measure, then the other must also be of positive measure. Hence, from the paragraph after (5.7) we obtain that

$$|l_k - l| \leq \frac{1}{k}.$$

Due to Corollary 5.1 applied to u and u_k, we have that for a.e. a for which $u(a) \neq u_k(a)$ we have either

$$\left(\ \lambda(a) > l \ \text{ and } \ \lambda_k(a) \leq l_k \leq l + \frac{1}{k}\right) \ \text{ or } \ \left(\ \lambda(a) < l \ \text{ and } \ \lambda_k(a) \geq l_k \geq l - \frac{1}{k}\right).$$

Having in mind that $\|\lambda_k - \lambda\|_{L_\infty(\Omega)} < 1/k$, we obtain that for every k

$$|\lambda(a) - l| \leq \frac{2}{k}$$

on a set of measure ε. In a standard way this implies that $\Omega_0(l)$ is of positive measure, which is a contradiction. □

COROLLARY 5.2. *Let λ be as in Lemma* 5.1. *Then problem* (5.1) *subject to* (3.4) *has a unique solution.*

Proof. It is enough to apply the above lemma for $\tilde{\lambda} = \lambda$. □

Example 1. We shall define an example that will provide several counterfacts showing that the regularity assumption is essential. Let $\omega > 1$, $\bar{u}(a) = 1$, $\beta = 1$, and $\mu(a) = 0$. Let us fix $\kappa \in (0,1)$ and $\theta = \omega - 1 + \kappa$. Define $p(a) = 0$ for $a \in [\theta, \omega]$. The value of $p(a)$ for $a < \theta$ will be defined below. For a given $\nu > 0$ denote by $\lambda[\nu]$ the solution of (5.3) on $[\theta, \omega]$. Moreover, define

$$\alpha = -\kappa\lambda[1](\theta) - \int_\theta^\omega \lambda[1](a)\,\mathrm{d}a, \quad p(a) = p^0 := r\lambda[1](\theta) + 1 \ \text{ for } \ a \in [0, \theta).$$

This completes the definition of the problem. One can check directly that the solution $\lambda[1](a)$ of (5.3) with $\kappa = 1$ is constant and equals $\lambda[1](\theta)$ on $[0, \theta]$. Let $u^* : [0, \theta] \mapsto [0, 1]$ be any measurable function such that $\int u^*(a)\,\mathrm{d}a = \kappa$. Define

$$u(a) = \begin{cases} u^*(a) & \text{for } a \in [0, \theta), \\ 1 & \text{for } a \in [\theta, \omega]. \end{cases} \tag{5.8}$$

Since $\kappa + (\omega - \theta) = 1$, u is an admissible control. Moreover, due to the time-invariance of $\lambda[1]$ on $[0, \theta]$ and the inequality $\lambda[1](\theta) < \lambda[1](a)$ for $a > \theta$, u satisfies (5.1). Equality

(5.2) is also satisfied due to the choice of α. Thus u satisfies the necessary condition formulated in Theorem 5.1.

Now we shall perturb the function p on $[0,\theta]$ in the following way: for a "small" (in absolute value) real number δ, and for a given value of the parameter $\nu > 0$ we define

$$p[\nu,\delta](a) = \nu p^0 + (1 + r(\theta - a))\delta, \quad a \in [0,\theta].$$

If we solve (5.3) for any ν and the above $p[\nu,\delta]$ we obtain for the solution $\lambda[\nu,\delta]$

$$\lambda[\nu,\delta](a) = \nu\lambda[1](a), \ \ t \in [\theta,\omega], \qquad \lambda[\nu,\delta](a) = \nu\lambda[1](a) + (\theta - a)\delta, \ \ a \in [0,\theta).$$

Hence, there is a unique maximizer $u[\nu,\delta]$ in (5.1), and in the case of $\delta > 0$ its structure is

$$u[\nu,\delta](a) = \begin{cases} 1 & \text{if } a \in [0, x(\nu,\delta)] \cup [y(\nu,\delta),\omega], \\ 0 & \text{elsewhere,} \end{cases}$$

while in the case $\delta < 0$ (in which case $\lambda[\nu,\delta]$ is monotone increasing) it is

$$u[\nu,\delta](a) = \begin{cases} 1 & \text{if } a \in [\omega - 1,\omega], \\ 0 & \text{elsewhere.} \end{cases}$$

The values $x(\nu,\delta)$ and $y(\nu,\delta)$ can be specified as the solutions of the system of transcendent equations

$$\begin{aligned} &\lambda[\nu,\delta](x) = \lambda[\nu,\delta](y), \\ &x + (\omega - y) = 1. \end{aligned}$$

The last system is rather easy to investigate and one can prove that $x(\nu,\delta) = \kappa + O(|\delta| + |\nu - 1|)$, $y(\nu,\delta) = \theta + O(|\delta| + |\nu - 1|)$. Then we chose ν from (5.2). This has a unique solution $\nu(\delta)$, and $\nu(\delta) \longrightarrow 1$ as $\delta \longrightarrow 0$. We skip the obvious details of the above construction.

Denote $u_\delta = u[\nu(\delta),\delta]$, $\delta \neq 0$. The problem with $\delta \neq 0$ satisfies the regularity assumption, and u_δ is by its construction the only admissible control satisfying the maximum principle (4.1)–(4.3) for the function $p_\delta = p[\nu(\delta),\delta](a)$. Hence u_δ is optimal. We have that p_δ converges uniformly to p^0 and u_δ converges in measure (and even stronger) to either

$$u^+(a) = \begin{cases} 1 & \text{if } a \in [0,\kappa] \cup [\theta,\omega], \\ 0 & \text{elsewhere,} \end{cases} \qquad \text{or} \qquad u^-(a) = \begin{cases} 1 & \text{if } a \in [\omega - 1,\omega], \\ 0 & \text{elsewhere} \end{cases}$$

(depending on whether $\delta > 0$ or $\delta < 0$). Then a standard upper semicontinuity argument for the solution set implies that both u^+ and u^- are optimal controls for the problem with $p = p^0$. So if the regularity assumption is violated we obtain the following.

Counterfact 1. The uniqueness claim in Theorem 5.1 is false.

Since each of u^- and u^+ is optimal independently of the initial state, then

$$u(t,a) = \begin{cases} u^-(a) & \text{for } t \in [2k\omega, (2k+1)\omega), \\ u^+(a) & \text{for } t \in [(2k+1)\omega, (2k+2)\omega), \ \ k = 0,1,\ldots \end{cases}$$

is also optimal. We obtain the following.

Counterfact 2. The time-invariance claim in Theorem 5.1 is false.

Counterfact 3. There is an optimal control u such that neither $u(t,\cdot)$ nor the corresponding trajectory $(M(t,\cdot),R(t))$ converges with $t \longrightarrow \infty$.

The question remains whether a convex combination of u^- and u^+ is also a solution. The answer need not be positive because the objective value $J(u)$ need not be a concave functional of the control u (see footnote 5). We investigated this question numerically and found that $J(0.5(u^- + u^+)) << J(u^-) = J(u^+)$.

Numerical "counterfact" 4. If the regularity assumption is violated, then the maximum principle is not a sufficient condition.

6. Strong ergodicity of the optimal solution. It is accepted in the population dynamics (see, e.g., [13, 4]) that *weak ergodicity* means that two populations with different initial age-distributions and the same (time-dependent) data become asymptotically identical (although a limit may fail to exist). *Strong ergodicity* means that the age-density of the population tends to a steady-state, which is independent of the initial density, if the data of the problem are convergent at infinity. In this section we obtain a strong ergodicity result for the optimal control of problem (1.1)–(1.5). Having this result, the issue of the asymptotic convergence of the corresponding optimal trajectory becomes a classical one (cf. [12, 20]).[8] The issue of weak ergodicity of the optimal solution is more complicated, requires additional conditions, and will not be discussed in this paper.

THEOREM 6.1 (strong ergodicity of the optimal control). *Let* $(\tilde{M}_0,\tilde{\mu},\tilde{p})$ *and* (M_0,μ,p) *be two triples of data for which the standing assumptions are fulfilled. Let* (μ,p) *be time-invariant (i.e., independent of* t*) and satisfy the regularity assumption. Assume, moreover, that*

$$\lim_{T\to\infty} \|(\tilde{\mu},\tilde{p}) - (\mu,p)\|_{L_\infty([T,\infty)\times\Omega)} = 0. \tag{6.1}$$

Then for any optimal control, $\tilde{u}$*, corresponding to* $(\tilde{M}_0,\tilde{\mu},\tilde{p})$*, and for the unique control,* u*, corresponding to* (M_0,μ,p)*, it holds that*

$$\lim_{t\to\infty} \operatorname{meas}\{a:\ \tilde{u}(t,a) \neq u(a)\} = 0.$$

Proof. As shown in section 4, the standing assumptions imply that the closed adjoint system (4.2), (4.3) for the problem with "˜" has a solution $\tilde{\xi}$ in $L_\infty(D)$. This solution satisfies (4.2), (4.3) also for the data (μ,p), with an additional term

$$\delta(t,a) = (\tilde{\mu}-\mu)\tilde{\xi} - (\tilde{\mu}-\mu)(\alpha+\sigma(\tilde{\xi})) - \beta(\tilde{p}-p).$$

Since $\|\delta\|_{L_\infty([T,\infty)\times\Omega)} = 0$ due to $\tilde{\xi} \in L_\infty(D)$ and (6.1), we obtain from Lemma 4.1 that

$$\lim_{T\to\infty} \|\tilde{\xi} - \xi\|_{L_\infty([T,\infty)\times\Omega)} = 0.$$

We know from Theorem 5.1 that ξ is time-invariant. In the proof of this theorem we have seen that λ in the formulation of that theorem differs from ξ by a constant. The

[8] For example, in the stationary case, and with the optimal $u(a)$ plugged into (1.1), equation (2.4) for $R(t)$ is a convolution renewal equation and its solution is convergent with $t \longrightarrow \infty$ according to [20, Theorem 7.1, Chapter 15]. We note that the convergence of R may fail in a discrete-time model with a stationary mortality rate and a stationary recruitment rule. A trivial example is the one of recruitment only at age zero. In our considerations such behavior is prohibited by the constraint $u(t,a) \le \bar{u}(a)$.

maximum principle (Theorem 3.1) claims that for a.e. t the functions $\tilde{u}(t,\cdot)$ and $u(\cdot)$ maximize

$$\int_\Omega \tilde{\xi}(t,a)v(a)\,\mathrm{d}a \quad \text{and} \quad \int_\Omega \xi(a)v(a)\,\mathrm{d}a,$$

respectively, subject to (3.4). Moreover, for every $\delta > 0$ we have

$$\|\tilde{\xi}(t,\cdot) - \xi(\cdot)\|_{L_\infty(\Omega)} < \delta$$

for a.e. sufficiently large t. Then Lemma 5.1 implies the claim of the theorem. □

Remark 6.1. Counterfact 4 in Example 1, section 5, shows that the strong ergodicity does not hold unless an appropriate additional assumption, such as the regularity assumption, is posed.

7. The principle of bipolar recruitment. In the previous section we proved that under the regularity assumption the optimal solution with time-dependent data approaches (in the sense specified in Theorem 6.1) the solution for the limit (time-invariant) data if the data (μ, p) converge with time (in the sense of Theorem 6.1). On the other hand, in section 5 we proved that the optimal control (i.e., the recruitment density) for time-invariant data is time-invariant and (thanks to this) is characterized in terms of an ODE (see (5.3)). Therefore, in this section we study in more detail the structure of the optimal control for stationary data, focusing on the case $p(a) = -a$. In this case the meaning of the problem (1.1)–(1.5) is that two objectives are to be optimized in the Pareto sense: the number of recruitments are to be high, and the average age of the organization is to be low. As suggested in [24] these two objectives are contradictory, which makes the problem interesting. The reason for which we focus the analysis on the average age is that this specific problem is under intensive discussion in many European academies of sciences due to the considerable aging of their members [24, 14, 15]. The analysis for other reasonable productivity functions, $p(a)$, such as a concave first-increasing-then-decreasing-with-age productivity (which is reasonable, according to [30]) gives also interesting qualitative results under specific joint conditions for μ and p, which we do not present here.

PROPOSITION 7.1 (bipolar recruitment principle). *Consider problem* (1.1)–(1.5) *with* $p(t,a) = -a$. *Assume that* $\mu(t,a) = \mu(a)$ *is a time-invariant, continuously differentiable, and nondecreasing function which equals zero on some interval* $[0, a_0)$ *and is strongly convex on* $(a_0, \omega]$. *Then the optimal recruitment density,* u, *is unique, time-invariant, and independent of the initial density* M_0 *and has the following structure. There are numbers* $0 \le \theta < \tau < \omega$ *such that*

$$u(a) = \begin{cases} \bar{u}(a) & \text{for } a \in [0,\theta) \cup (\tau,\omega], \\ 0 & \text{for } a \in [\theta,\tau]. \end{cases} \tag{7.1}$$

Remark 7.1. The natural mortality rate in the ages above 30 satisfies the assumptions for μ. We stress also the remarkable fact that the second interval of recruitment is always nondegenerate: $\tau < \omega$.

Proof. We apply Theorem 5.1 and characterization (5.6) of the optimal control. The regularity assumption is apparently fulfilled, since the function $\mu(a) + dp(a) = e$ may have at most two zeros.

Since λ is twice continuously differentiable, all we have to prove is that λ has no local maxima in $(0, \omega)$.

Assume that $a \in (0, \omega)$ is a local maximizer of λ, and $\lambda(a) \geq 0$. Then

$$\lambda''(a) = \mu'(a)\lambda(a) + \beta > 0,$$

which is a contradiction. Now assume that $\lambda(a) < 0$. Since $\lambda(\omega) = 0$, there must be a local minimizer $b \in [a, \omega)$ with $\lambda(b) \leq \lambda(a)$. Then

$$0 \geq \lambda''(a) = \mu'(a)\lambda(a) + \beta \geq \mu'(b)\lambda(a) + \beta \geq \mu'(b)\lambda(b) + \beta = \lambda''(b) \geq 0.$$

Then all the inequalities must be equalities. Since $\lambda(a) < 0$, this implies $\mu'(a) = \mu'(b)$, which yields $a \in (0, a_0]$. Since μ is identically zero on $[0, a_0)$, we have $\lambda''(a) = \beta$, which is a contradiction. Thus the optimal $u(a)$ has the structure (7.1). It remains only to prove that $\tau < \omega$. If this is not the case, then $\lambda(a) \geq 0$ for all a, for which $u(a) > 0$. This contradicts (5.2). □

Now we give an intuitive argument showing that the additional assumptions in the above theorem are essential (although not necessary) for the obtained bipolar recruitment principle. If μ is a bounded approximation of the δ-function concentrated at $\omega/2$, then the age $\omega/2$ would appear as a premature retirement age, similarly to ω. Then by the same argument as in Proposition 7.1 $u(t, a)$ would be positive shortly before $\omega/2$. It would be positive also before ω (since μ is bounded). If ω is sufficiently large, then three disjoint intervals with $u(t, a) > 0$ would appear.

We stress that in practical applications the principle of bipolar recruitment should not be taken in an absolute sense, as far as usually many other criteria (different than the recruitment intensity and the average age of the members of the organization) are taken into account. The essence of the result is that if the last mentioned two criteria matter for the organization (which is, indeed, the case for many organizations such as academies of sciences or academies of awards), then they have a polarizing effect on the optimal recruitment policy: they shift the recruitment partly to younger and partly to older ages, decreasing in this way the middle-age recruitment. An interesting interpretation of this result is given by Warren Sanderson.[9] The essence is that an academy of awards should focus on awarding relatively young talents for recent outstanding achievements and, on the other hand, old persons for their life-long contributions. Detailed policy-oriented considerations and interpretations are given in the forthcoming paper [14].

Appendix. First we prove Proposition 2.1.

Proof. The proof uses the idea from [3]. This idea is substantially modified to fit to our problem; therefore, we present a detailed proof, as required by one of the referees.

As before we use the notation $D = [0, \infty) \times \Omega$ and also $D_T = [0, T] \times \Omega$. Consider a maximizing sequence $\{u_k\}$ of admissible controls such that

$$J(u_k) \geq J^* - \frac{1}{k}, \tag{A.1}$$

where $J(u_k)$ is the objective value for u_k, together with the corresponding solution (M_k, R_k) of (1.1)–(1.3), and J^* is the supremum of the objective function in the set of admissible controls. According to Lemma 2.1 the sequence $\{M_k\}$ is well defined in D, and the functions M_k are bounded uniformly in k. Then the sequence $e^{-rt}M_k \in L_1(D)$

[9] Warren Sanderson, Professor of Economics, SUNY-Stony Brook, Stony Brook, NY 11794-4384, USA. Personal communications with the authors.

is weakly relatively compact due to the Dunford–Pettis criterion.[10] Therefore, there exists a subsequence, denoted also by M_k, such that

$$e^{-rt}M_k \longrightarrow e^{-rt}M_0 \quad L_1(D)\text{-weakly.}$$

Clearly M_0 is bounded by the same constant as M_k. According to the Mazur theorem there exists a sequence

$$e^{-rt}\tilde{M}_k = \sum_{i=k}^{n_k} p_i^k e^{-rt}M_i, \quad p_i^k \geq 0, \quad \sum_{i=k}^{n_k} p_i^k = 1,$$

convergent to $e^{-rt}M_0$ in $L_1(D)$. Clearly $\tilde{M}_k$ are uniformly bounded, and $\tilde{M}_k \longrightarrow M_0$ in $L_1(D_T)$ for every $T > 0$. Define

$$\tilde{R}_k(t) = \sum_{i=k}^{n_k} p_i^k R_i(t) = \tilde{M}_k(t,\omega) + \int_\Omega \mu(t,a)\tilde{M}_k(t,a)\,\mathrm{d}a. \tag{A.2}$$

Let $\tilde{u}_0 : \Omega \mapsto \mathbf{R}$ be an arbitrary measurable function satisfying the control constraints (1.4). Let us define

$$\tilde{u}_k(t,a) = \begin{cases} \sum_{i=k}^{n_k} \frac{p_i^k R_i(t) u_i(t,a)}{\tilde{R}_k(t)} & \text{if } \tilde{R}_k(t) > 0, \\ \tilde{u}_0(a) & \text{elsewhere.} \end{cases}$$

Obviously $\tilde{u}_k$ satisfies (1.4); therefore, it is an admissible control. We have

$$\frac{\partial \tilde{M}_k}{\partial t} + \frac{\partial \tilde{M}_k}{\partial a} = \sum_{i=k}^{n_k} p_i^k(-\mu M_i + R_i u_i) = -\mu\tilde{M}_k + \sum_{i=k}^{n_k} p_i^k R_i u_i = -\mu\tilde{M}_k + \tilde{R}_k\tilde{u}_k.$$

Thus $(\tilde{u}_k, \tilde{M}_k, \tilde{R}_k)$ is an admissible control-trajectory triple, for which $\tilde{M}_k \longrightarrow M_0$ in $L_1(D_T)$ and $e^{-rt}\tilde{M}_k \longrightarrow e^{-rt}M_0$ in $L_1(D)$. Moreover, passing to a subsequence, we may assume that $\tilde{M}_k$ converges to M_0 almost everywhere and that $e^{-rt}\tilde{u}_k$ converges to some $e^{-rt}u_0$ weakly in $L_1(D)$. In the next paragraph we prove that u_0 is an admissible control.

For every measurable and bounded set $\Gamma \subset [0,\infty)$ we have that

$$\int_\Gamma \int_\Omega \tilde{u}_k(t,a)\,\mathrm{d}a\,\mathrm{d}t \longrightarrow \int_\Gamma \int_\Omega u_0(t,a)\,\mathrm{d}a\,\mathrm{d}t.$$

Since $\tilde{u}_k$ satisfies (1.4), the left-hand side equals $\mathrm{meas}(\Gamma)$; hence

$$\int_\Gamma \int_\Omega u_0(t,a)\,\mathrm{d}a\,\mathrm{d}t = \mathrm{meas}(\Gamma).$$

Since Γ is an arbitrary measurable and bounded set, this implies that u_0 satisfies the equality in (1.4) for a.e. t. The inequality in (1.4) is obviously also satisfied.

Next we prove that the sequence $\{e^{-rt}\tilde{R}_k\}$ is convergent in $L_1(0,\infty)$. According to the definition of a solution and Lemma 2.1 on a.e. characteristic line $\{(t-s,\omega-s)\}_{s\in[0,\min\{t,\omega\}]}$ the functions $\tilde{M}_k$ are Lipschitz continuous uniformly in k and t. On

[10] One can work in the space L_1 weighted by the factor e^{-rt}, but we prefer to keep this factor explicit.

the other hand, $\tilde{M}_k(t-\cdot,\omega-\cdot)$ converges pointwise to $M_0(t-\cdot,\omega-\cdot)$ for a.e. t. Then $\tilde{M}_k(t-\cdot,\omega-\cdot)$ converges to $M_0(t-\cdot,\omega-\cdot)$ uniformly; hence the latter function is Lipschitz with the same constant as $\tilde{M}_k(t-\cdot,\omega-\cdot)$. Then $M_0(t,\omega)$ is well defined in $L_\infty(0,\infty)$. Moreover, $\tilde{M}_k(\cdot,\omega)$ converges to $M_0(\cdot,\omega)$ almost everywhere; hence $e^{-rt}\tilde{M}_k(\cdot,\omega) \longrightarrow e^{-rt}M_0(\cdot,\omega)$ in $L_1(0,\infty)$. Due to (A.2), the sequence $\tilde{R}_k(t)$ converges almost everywhere to

$$R_0(t) = M_0(t,\omega) + \int_0^\omega \mu(t,a)M_0(t,a)\,\mathrm{d}a;$$

hence $e^{-rt}\tilde{R}_k$ converges to $e^{-rt}R_0$ in $L_1(0,\infty)$. To verify that (u_0, M_0, R_0) satisfies (1.1) we integrate in a the representation (2.2) for the solution triple $(\tilde{u}_k, \tilde{M}_k, \tilde{R}_k)$ on a measurable set $\Gamma \subset \Omega$. Due to the established properties of this sequence we may pass to the limit, (u_0, M_0, R_0). Since Γ is arbitrary, we obtain that (u_0, M_0, R_0) satisfies (2.2) for a.e. (t,a); hence it is a solution of (1.1).

From the convexity of A with respect to (M,w) we have

$$\begin{aligned}
J(\tilde{u}_k) &= \int_0^\infty e^{-rt}\left[\alpha\tilde{R}_k(t) + \beta\int_\Omega A(t,a,\tilde{M}_k(t,a),\tilde{R}_k(t)\tilde{u}_k(t,a))\,\mathrm{d}a\right]\mathrm{d}t\\
&= \int_0^\infty e^{-rt}\Bigg[\alpha\sum_{i=k}^{n_k} p_i^k R_i(t)\\
&\qquad + \beta\int_\Omega A\Bigg(t,a,\sum_{i=k}^{n_k} p_i^k M_i(t,a),\sum_{i=k}^{n_k} p_i^k R_i(t)u_i(t,a)\Bigg)\mathrm{d}a\Bigg]\mathrm{d}t\\
&\geq \sum_{i=k}^{n_k} p_i^k J(u_i) \geq \sum_{i=k}^{n_k} p_i^k\left(J^* - \frac{1}{i}\right) \geq J^* - \frac{1}{k}.
\end{aligned}$$

Finally, from the upper semicontinuity of the objective function with respect to u in the $L_1(T)$-weak topology (due to the concavity of A in w), the Lipschitz continuity of A in (M,w), the L_1-convergence of $\tilde{M}_k$ and $\tilde{R}_k$, and the boundedness of A in the domain of integration below, we obtain for every $T>0$

$$\begin{aligned}
J^* &\leq \limsup_k\left(J(\tilde{u}_k) + \frac{1}{k}\right) = \limsup_k J(\tilde{u}_k)\\
&= \limsup_k \int_0^\infty e^{-rt}\left[\alpha\tilde{R}_k(t) + \beta\int_\Omega A(t,a,\tilde{M}_k(t,a),\tilde{R}_k(t)\tilde{u}_k(t,a))\,\mathrm{d}a\right]\mathrm{d}t\\
&= \int_0^\infty e^{-rt}\alpha R_0(t) + \beta\limsup_k\int_0^\infty\int_\Omega e^{-rt}A(t,a,M_0(t,a),R_0(t)\tilde{u}_k(t,a))\,\mathrm{d}a\,\mathrm{d}t\\
&\leq \int_0^\infty e^{-rt}\alpha R_0(t) + \beta\limsup_k\int_0^T\int_\Omega e^{-rt}A(t,a,M_0(t,a),R_0(t)\tilde{u}_k(t,a))\,\mathrm{d}a\,\mathrm{d}t\\
&\qquad + \frac{\beta}{r}e^{-rT}C\\
&\leq \int_0^\infty e^{-rt}\alpha R_0(t) + \beta\int_0^T\int_\Omega e^{-rt}A(t,a,M_0(t,a),R_0(t)u_0(t,a))\,\mathrm{d}a\,\mathrm{d}t + \frac{\beta}{r}e^{-rT}C\\
&\leq J(u_0) + \frac{2\beta}{r}e^{-rT}C.
\end{aligned}$$

Since T is an arbitrary positive number, we obtain that $J^* \leq J(u_0)$; hence u_0 is an optimal control. □

Below we prove Lemma 3.1.

Proof. The proof is split into several steps.

Step 1. Let us fix a function $\eta \in L_\infty(0,\infty)$ and a positive number T and consider (3.1) in the domain $D_T = [0,T] \times \Omega$, with the side conditions (3.3) and

$$\xi(T,a) = 0, \quad a \in \Omega. \tag{A.3}$$

This equation is linear and can be solved along the characteristics, which results in the following explicit formula:

$$\begin{aligned}\xi(t,a) &= \varphi(t,a,t+\omega-a)\chi_{[0,T]}(t+\omega-a)\eta(t+\omega-a)\\ &\quad + \int_t^T \varphi(t,a,\tau)\chi_\Omega(a-t+\tau)[\mu(\tau,a-t+\tau)\eta(\tau) - \beta A_M(\tau,a-t+\tau)]\,\mathrm{d}\tau,\end{aligned} \tag{A.4}$$

where here and below the argument (t,a) of the function A_M is a substitution for $(t,a,M(t,a),u(t,a))$, χ_S is the indicator function of the set S, and

$$\varphi(t,a,\tau) = e^{-\int_t^\tau (r+\mu(\theta,a-t+\theta))\,\mathrm{d}\theta}. \tag{A.5}$$

We shall determine $\eta(t) = \eta_T(t)$ in such a way that (3.2) is also fulfilled. Due to formula (A.4) for ξ, (3.2) becomes

$$\begin{aligned}\eta(t) &= \alpha + \int_\Omega \varphi(t,a,t+\omega-a)\chi_{[0,T]}(t+\omega-a)\eta(t+\omega-a)u(t,a)\,\mathrm{d}a\\ &+ \int_\Omega\int_t^T \varphi(t,a,\tau)\chi_\Omega(a-t+\tau)[\mu(\tau,a-t+\tau)\eta(\tau) - \beta A_M(\tau,a-t+\tau)]u(t,a)\,\mathrm{d}\tau\,\mathrm{d}a\\ &+\beta\int_\Omega A_w(t,a)u(t,a)\,\mathrm{d}a.\end{aligned}$$

Changing the variable $t+\omega-a=\tau$ in the first integral and changing the order of integration in the second one, we obtain that η has to satisfy the integral equation

$$\eta(t) = f_T(t) + \int_t^T K(t,\tau)\eta(\tau)\,\mathrm{d}\tau, \tag{A.6}$$

where

$$\begin{aligned}f_T(t) &= \alpha - \beta\int_t^T\int_\Omega \varphi(t,a,\tau)\chi_\Omega(a-t+\tau)A_M(\tau,a-t+\tau)u(t,a)\,\mathrm{d}a\,\mathrm{d}\tau\\ &\quad +\beta\int_\Omega A_w(t,a)u(t,a)\,\mathrm{d}a,\\ K(t,\tau) &= \varphi(t,t+\omega-\tau,\tau)\chi_{[0,t+\omega]}(\tau)u(t,t+\omega-\tau)\\ &\quad + \int_\Omega \varphi(t,a,\tau)\chi_\Omega(a-t+\tau)\mu(\tau,a-t+\tau)u(t,a)\,\mathrm{d}a.\end{aligned}$$

Clearly, (A.6) is a Volterra equation of the second kind (inverse in time); therefore, it has a unique solution $\eta_T \in L_\infty(0,T)$.

Step 2. Next we shall prove uniform boundedness of $\eta_T(t)$ when $T \longrightarrow \infty$. First, since $\varphi(t,a,\tau) \le 1$ and u satisfies (1.4), we have

$$|f_T(t)| \le \alpha + \beta\bar{A}\int_t^{T(t)}\int_0^{t+\omega-\tau} u(t,a)\,\mathrm{d}a\,\mathrm{d}\tau + \beta\bar{A} \le \alpha + \beta\omega\bar{A} + \beta\bar{A} =: c_0, \tag{A.7}$$

where $T(t) = \min\{T, t+\omega\}$ and $\bar{A}$ is a bound for the derivatives A_M and A_w. Moreover, obviously

$$K(t,\tau) \le \bar{u} + \omega\bar{\mu} =: c_1. \tag{A.8}$$

Define the operator $\mathcal{F}_T : L_\infty(0,T) \mapsto L_\infty(0,T)$ by

$$\mathcal{F}_T(\eta)(t) = \int_t^T K(t,\tau)\eta(\tau)\,\mathrm{d}\tau.$$

Since K is nonnegative we have

$$\|\mathcal{F}_T(\eta)\|_\infty \le \operatorname*{ess\,sup}_{t\in[0,T]} \int_t^T K(t,\tau)\,\mathrm{d}\tau \|\eta\|_\infty. \tag{A.9}$$

For $t \le T-\omega$ we have

$$\begin{aligned}\int_t^T K(t,\tau)\,\mathrm{d}\tau = &\int_t^{t+\omega} \varphi(t,t+\omega-\tau,\tau)u(t,t+\omega-\tau)\,\mathrm{d}\tau \\ &+ \int_t^{t+\omega}\int_0^{t+\omega-\tau} \varphi(t,a,\tau)\mu(\tau,a-t+\tau)u(t,a)\,\mathrm{d}a\,\mathrm{d}\tau.\end{aligned}$$

Consider the expression

$$\begin{aligned}G(t) &= \int_t^{t+\omega} \frac{\partial}{\partial\tau}\int_0^{t+\omega-\tau} \varphi(t,a,\tau)u(t,a)\,\mathrm{d}a\,\mathrm{d}\tau \\ &= -\int_t^{t+\omega}\left[\varphi(t,t+\omega-\tau,\tau)u(t,t+\omega-\tau) - \int_0^{t+\omega-\tau}\frac{\partial\varphi}{\partial\tau}(t,a,\tau)u(t,a)\,\mathrm{d}a\right]\mathrm{d}\tau \\ &= -\int_t^{t+\omega}\Bigg[\varphi(t,t+\omega-\tau,\tau)u(t,t+\omega-\tau) \\ &\qquad + \int_0^{t+\omega-\tau}\varphi(t,a,\tau)(r+\mu(\tau,a-t+\tau))u(t,a)\,\mathrm{d}a\Bigg]\mathrm{d}\tau.\end{aligned}$$

Hence,

$$\int_t^T K(t,\tau)\,\mathrm{d}\tau = -G(t) - r\int_t^{t+\omega}\int_0^{t+\omega-\tau}\varphi(t,a,\tau)u(t,a)\,\mathrm{d}a\,\mathrm{d}\tau.$$

On the other hand, having in mind (1.4) and that $\varphi(t,a,t) = 1$ we obtain that

$$G(t) := \int_0^{t+\omega-\tau}\varphi(t,a,\tau)u(t,a)\,\mathrm{d}a\ \Big|_t^{t+\omega} = -1.$$

Thus

$$\int_t^T K(t,\tau)\,\mathrm{d}\tau \le 1 - r\int_t^{t+\omega}\int_0^{t+\omega-\tau}\varphi(t,a,\tau)u(t,a)\,\mathrm{d}a\,\mathrm{d}\tau.$$

We estimate

$$\begin{aligned}&r\int_t^{t+\omega}\int_0^{t+\omega-\tau}\varphi(t,a,\tau)u(t,a)\,\mathrm{d}a\,\mathrm{d}\tau \ge r\int_t^{t+\omega}\int_0^{t+\omega-\tau} e^{-(\tau-t)(r+\bar{\mu})}u(t,a)\,\mathrm{d}a\,\mathrm{d}\tau \\ &\ge re^{-\omega(r+\bar{\mu})}\int_0^\omega\int_t^{t+\omega-a}\mathrm{d}\tau u(t,a)\,\mathrm{d}a = re^{-\omega(r+\bar{\mu})}\int_0^\omega(\omega-a)u(t,a)\,\mathrm{d}a.\end{aligned}$$

The value of the last integral is at least

$$d := \int_{\omega-s}^{\omega} (\omega - a)\bar{u}(a)\,\mathrm{d}a > 0,$$

where s is determined from the equality

$$\int_{\omega-s}^{\omega} \bar{u}(a)\,\mathrm{d}a = 1.$$

Thus for $t \le T - \omega$

$$\int_t^T K(t,\tau)\,\mathrm{d}\tau \le 1 - \gamma, \tag{A.10}$$

where

$$0 < \gamma := re^{-\omega(r+\bar{\mu})}d.$$

Due to (A.7) and (A.8) it is standard to prove that the solution of the Volterra equation (A.6) is bounded on the interval $[T-\omega, T]$, uniformly in T, by a constant c_2. Combining this with (A.10) and using (A.9) we obtain

$$|\eta_T(t)| \le c_0 + \|\mathcal{F}_T(\eta)\| \le c_0 + \max\{(1-\gamma)\|\eta_T\|, c_1c_2\}.$$

Hence,

$$|\eta_T(t)| \le \max\left\{\frac{c_0}{\gamma}, c_0 + c_1c_2\right\}.$$

This proves the uniform boundedness of η_T in L_∞.

Step 3. Now we prove the following property: for every $\varepsilon > 0$ and $T > 0$ there exists $\bar{T} > T$ such that for every $T', T'' \ge \bar{T}$ it holds that

$$\|\eta_{T'} - \eta_{T''}\|_{L_\infty([0,T])} \le \varepsilon. \tag{A.11}$$

Let us take $T'' > T' \ge \bar{T}$, where $\bar{T}$ will be defined later, assuming now only that $\bar{T} \ge T + \omega$.

Denote $\Delta(t) = \eta_{T''}(t) - \eta_{T'}(t)$. For $t \le \bar{T} - \omega$ it holds that $f_{T'}(t) = f_{T''}(t)$ and for $\tau > t + \omega$ it holds that $K(t,\tau) = 0$. Therefore, for $t \le \bar{T} - \omega$ we have that

$$\begin{aligned}\Delta(t) &= f_{T''}(t) - f_{T'}(t) + \int_t^{T''} K(t,\tau)\eta_{T''}(\tau)\,\mathrm{d}\tau - \int_t^{T'} K(t,\tau)\eta_{T'}(\tau)\,\mathrm{d}\tau \\ &= \int_t^{t+\omega} K(t,\tau)\Delta(\tau)\,\mathrm{d}\tau.\end{aligned}$$

Denote $t_k = \bar{T} - k\omega$, $k = 0, \dots,$ and for the first $k = \bar{k}$ for which $t_k \le 0$ we redefine $t_{\bar{k}} = 0$. Then consider

$$\Delta_k = \operatorname{ess\,sup}\{\Delta(t) :\ t \in [t_{k+1}, t_k]\}, \quad k = 0, \dots, \bar{k}.$$

From Step 2 of the proof we know that $\Delta_0 \le c$, where c is independent of $T, \bar{T}, T', T''$. Moreover,

$$\begin{aligned}\Delta_k &= \operatorname*{ess\,sup}_{t\in[t_{k+1},t_k]} \Delta(t) \le \operatorname*{ess\,sup}_{t\in[t_{k+1},t_k]} \int_t^{t_k} K(t,\tau)\Delta(\tau)\,\mathrm{d}\tau + \int_{t_k}^{t+\omega} K(t,\tau)\Delta(\tau)\,\mathrm{d}\tau \\ &\le \max\{\Delta_k, \Delta_{k-1}\} \int_t^{t+\omega} K(t,\tau)\,\mathrm{d}\tau \le (1-\gamma)\max\{\Delta_k, \Delta_{k-1}\},\end{aligned}$$

which implies that

$$\Delta_k \le (1-\gamma)\Delta_{k-1} \le \cdots \le (1-\gamma)^k c.$$

Now, given $\varepsilon > 0$ and T, we choose $\bar{T}$ in such a way that for $k > (\bar{T} - T)/\omega$ it holds that $(1-\gamma)^k c < \varepsilon$. Then we have

$$|\Delta(t)| \le \varepsilon \quad \text{for} \quad t \le T.$$

Step 4. The uniform boundedness of η_T and (A.4) imply that the solution ξ_T of (3.1) on D_T with $\eta = \eta_T$ is also bounded uniformly in T.

Now we shall prove that ξ_T has the same property as that established in Step 5 for η_T: for every $\varepsilon > 0$ and $T > 0$ there exists $\bar{T} > T$ such that for every $T', T'' \ge \bar{T}$ it holds that

$$\|\xi_{T'} - \xi_{T''}\|_{L_\infty(D_T)} \le \varepsilon. \tag{A.12}$$

Let $\bar{T}$, T', and T'' be chosen as in Step 3, but for the numbers $T+\omega$ and $\varepsilon/(1+\omega\bar{\mu})$ (instead of T and ε). Then we have

$$|\Delta(t)| = |\eta_{T''}(t) - \eta_{T'}(t)| \le \frac{\varepsilon}{(1+\omega\bar{\mu})} \quad \text{for} \ \ t \le T+\omega.$$

Using (A.4) we obtain that for $t \le T$ and $a \in [0, \omega]$

$$|\xi_{T''}(t,a) - \xi_{T'}(t,a)| \le \left(1 + \int_t^{t+\omega} \bar{\mu}\, \mathrm{d}\tau\right) \|\Delta\|_{L_\infty(0,T+\omega)} \le \varepsilon.$$

Step 5. The properties of η_T and ξ_T proven in steps 3 and 4 imply that the sequence of restrictions $\{(\xi_{T_N}, \eta_{T_N})_{|[0,T]}\}_N$ (with $T_N \to +\infty$) is fundamental. From the completeness of L_∞ the limit exists, and letting $T \to \infty$, we extend it to $[0, \infty)$ (notice that the extension remains in L_∞ due to the uniform boundedness of ξ_{T_N} and η_{T_N}). For this limit function $(\xi, \eta) \in L_\infty$ it holds that for every $\varepsilon > 0$ and every $T > 0$ one can find $\bar{T} > T$ such that for every $\tilde{T} \ge \bar{T}$

$$\|\eta - \eta_{\tilde{T}}\|_{L_\infty([0,T])} + \|\xi - \xi_{\tilde{T}}\|_{L_\infty(D_T)} \le \varepsilon.$$

From here it easily follows that ξ is a solution of (3.1) in D, and η satisfies (3.2) almost everywhere in $[0, \infty)$.

Step 6. It remains to prove uniqueness of the solution. We skip this proof since it is identical to the proof of uniqueness for the closed adjoint system given in Corollary 4.1. ∎

REFERENCES

[1] A. A. Agrachev, G. Stefani, and P. Zezza, *Strong optimality for a bang-bang trajectory*, SIAM J. Control Optim., 41 (2002), pp. 991–1014.

[2] S. Aniţa, *Analysis and Control of Age-Dependent Population Dynamics,* Math. Model. Theory Appl., 11, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000.

[3] S. Aniţa, M. Iannelli, M.-Y. Kim, and E.-J. Park, *Optimal harvesting for periodic age-dependent population dynamics*, SIAM J. Appl. Math., 58 (1998), pp. 1648–1666.

[4] W. B. Arthur, *The ergodic theorems of demography: A simple proof*, Demography, 19 (1982), pp. 439–445.

[5] W. B. Arthur and T. J. Espenshade, *Immigration policy immigrants ages*, Population and Development Review, 14 (1988), pp. 315–326.

[6] S. M. Aseev and A. V. Kryazhimskiy, *The Pontryagin maximum principle and transversality conditions for a class of optimal control problems with infinite time horizons*, SIAM J. Control Optim., 43 (2004), pp. 1094–1119.
[7] J.-P. Aubin, N. Bonneuil, and F. Maurin, *Nonlinear structured population dynamics with co-variates*, Math. Population Studies, 9 (2000), pp. 1–31.
[8] V. Barbu, E. N. Barron, and R. Jensen, *The necessary conditions for optimal control in Hilbert spaces*, J. Math. Anal. Appl., 133 (1988), pp. 151–162.
[9] E. N. Barron and R. Jensen, *The Pontryagin maximum principle from dynamic programming and viscosity solutions to first-order partial differential equations*, Trans. Amer. Math. Soc., 298 (1986), pp. 635–641.
[10] E. Barucci and F. Gozzi, *Technology adoption and accumulation in a vintage capital model*, J. Economics, 74 (2001), pp. 1–38.
[11] M. Brokate, *Pontryagin's principle for control problems in age-dependent population dynamics*, J. Math. Biol., 23 (1985), pp. 75–101.
[12] T. A. Burton, *Volterra Integral and Differential Equations*, 2nd ed., Math. Sci. Engrg. 202, Elsevier, Amsterdam, 2005.
[13] J. E. Cohen, *Ergodic theorems in demography*, Bull. Amer. Math. Soc., 1 (1979), pp. 275–295.
[14] H. Dawid, G. Feichtinger, J. Goldstein, and V. M. Veliov, *Keeping a learned society young*, Research report of Institute of Mathematic Methods in Economics, Vienna University of Technologies, 2006 (submitted).
[15] G. Feichtinger, I. Freund, A. Prskawetz, V. M. Veliov, and M. Winkler-Dworak, *Age Dynamics and Optimal Recruitment Policies of Constant Sized Organizations*, An Application to the Austrian Academy of Sciences, 2007 (submitted).
[16] G. Feichtinger and A. Mehlmann, *The recruitment trajectory corresponding to particular stock sequences in Markovian person-flow models*, Math. Oper. Res., 1 (1976), pp. 175–184.
[17] G. Feichtinger and G. Steinmann, *Immigration into a population with fertility below replacement level—the case of Germany*, Population Studies, 46 (1992), pp. 275–284.
[18] G. Feichtinger, G. Tragler, and V. M. Veliov, *Optimality conditions for age-structured control systems*, J. Math. Anal. Appl., 288 (2003), pp. 47–68.
[19] U. Felgenhauer, *On stability of bang-bang type controls*, SIAM J. Control Optim., 41 (2003), pp. 1843–1867.
[20] G. Gripenberg, S.-O. Londen, and O. Staffans, *Volterra Integral and Functional Equations*, Cambridge University Press, Cambridge, UK, 1990.
[21] A. Ioffe and V. Tikhomirov, *Theory of Extremal Problems*, Nauka, Moscow, 1974 (in Russian).
[22] A. Ioffe and V. Tikhomirov, *Theory of Extremal Problems*, Stud. Math. Appl. 6, North–Holland, Amsterdam, 1979.
[23] H. Maurer and N. P. Osmolovskii, *Second order sufficient conditions for time-optimal bang-bang control*, SIAM J. Control Optim., 42 (2004), pp. 2239–2263.
[24] H. Leridon, *The demography of a learned society*, Population-E, 59 (2004), pp. 81–114.
[25] Zh. Luo, Ze-R. He, and W.-T. Li, *Optimal birth control for an age-dependent n-dimensional food chain model*, J. Math. Anal. Appl., 287 (2003), pp. 557–576.
[26] A. C. McKendrick, *Applications of mathematics to medical problems*, Proc. Edinburgh Math. Soc., 444 (1926), pp. 98–130.
[27] S. Mitra, *Generalization of the immigration and the stable population model*, Demography, 20 (1983), pp. 111–115.
[28] S. Preston, *The birth trajectory corresponding to particular sequences*, Theoretical Population Biology, 1 (1970), pp. 346–351.
[29] C. P. Schmertmann, *Immigrants' ages and the structure of stationary populations with below-replacement fertility*, Demography, 29 (1992), pp. 595–612.
[30] V. Skirbekk, *Age and individual productivity: A literature survey*, Vienna Yearbook of Population Research 2004, Austrian Academy of the Sciences Press, Vienna, Austria, 2004, pp. 133–153.
[31] V. M. Veliov, *On the bang-bang principle for linear control systems*, C. R. Acad. Bulgare Sci., 40 (1987), pp. 31–33.
[32] V. M. Veliov, *Error analysis of discrete approximations to bang-bang optimal control problems: The linear case*, Control Cybernet., 34 (2005), pp. 967–982.
[33] G. F. Webb, *Theory of Nonlinear Age-Dependent Population Dynamics*, Marcel Dekker, New York, 1985.

SIAM J. OPTIM.
Vol. 18, No. 3, pp. 1004–1026

STABILITY AND CONVERGENCE OF EULER'S METHOD FOR STATE-CONSTRAINED DIFFERENTIAL INCLUSIONS*

ROBERT BAIER†, ILYES AÏSSA CHAHMA‡, AND FRANK LEMPIO†

Abstract. A discrete stability theorem for set-valued Euler's method with state constraints is proved. This theorem is combined with known stability results for differential inclusions with so-called smooth state constraints. As a consequence, order of convergence equal to 1 is proved for set-valued Euler's method, applied to state-constrained differential inclusions.

Key words. Filippov theorem, set-valued Euler's method, differential inclusions with state constraints, stability and convergence of discrete approximations

AMS subject classifications. 49J24, 65L20, 34K28, 34A60

DOI. 10.1137/060661867

1. Introduction and preliminaries.

Differential inclusions appear in various fields of applications, e.g., in the study of (deterministic) perturbations of differential equations, in dynamical systems with discontinuous system equations, optimal control problems, viability theory, and especially climate impact research; cf., e.g., [2, 3, 14, 10, 1, 6].

An important subclass consists of differential inclusions with additional monotonicity properties which, in general, guarantee uniqueness of the solution of the initial value problem (cf., e.g., [2, 3, 4, 5, 20, 21]). Differential inclusions with Lipschitz right-hand sides (with respect to Hausdorff distance) in the usual sense form another important subclass. This latter subclass is the principal focus of this paper, which deals with stability and convergence properties of set-valued Euler's method for differential inclusions with state constraints.

The main result of this paper is the proof of a discrete stability theorem for a difference inclusion with state constraints in section 3, which serves as a basis for the convergence analysis for set-valued Euler's method in section 4. Intrinsically, this result is a variant of the Gronwall–Filippov–Wazewski theorem and, in fact, an existence theorem as well. Whereas the proofs for explicit difference inclusions with appropriate Lipschitz properties offer no difficulties, additional state constraints cause essential problems.

Fortunately, remarkable stability results for state-constrained differential inclusions have become available in the literature; cf. [22, 15, 17, 18, 7, 8, 23]. But discrete analogues for the approximation of all feasible trajectories under comparably weak conditions are still missing. Therefore, we concentrate on the so-called smooth case, where the state constraint is described by a single scalar inequality, resp., by a smooth signed distance function. This case has already been treated in [6], but contrary to [6] we allow time-dependent state constraints and improve the final error estimate.

In section 3 we give a rather complete analysis of the discrete situation, which heavily relies on the proof strategy in [15, Theorem 4.1] for the continuous problem.

*Received by the editors June 2, 2006; accepted for publication (in revised form) May 25, 2007; published electronically October 4, 2007.
http://www.siam.org/journals/siopt/18-3/66186.html
†Chair of Applied Mathematics, University of Bayreuth, D-95440 Bayreuth, Germany (robert.baier@uni-bayreuth.de, frank.lempio@uni-bayreuth.de).
‡Banco Cetelem, S.A., C/ Retama, 3, 3ª Planta, 28045 Madrid, Spain (aissa.chahma@cetelem.es).

In some respects, the discrete analysis is rather technical, and some additional difficulties have to be overcome. In particular, a discrete solution might not hit exactly the boundary of the state constraints, neighboring continuous solutions of feasible discrete solutions could violate the state constraints outside the grid, and consequently additional error terms appear in Taylor expansions.

However, we want to urgently emphasize the fact that only both stability results, the continuous *and* the discrete one together, will give us convergence results for discrete approximations of state-constrained differential inclusions. This is the essential subject of section 4, where order of convergence $\mathcal{O}(h)$ with respect to the step-size h is proved for set-valued Euler's method in the presence of state constraints.

In section 5, the results are applied to a differential inclusion resulting from a state-constrained bilinear control problem, which originally served as an academic test example for unconstrained problems and was communicated to us by Petar Kenderov. The order of convergence of the reachable sets of Euler's difference inclusion with state constraints to the corresponding reachable sets of the differential inclusion is visualized by computer tests. For a more detailed discussion and applications to climate impact research, see [6].

Hence, the main objective of this paper is the discrete approximation of the *whole* solution set of state-constrained differential inclusions, especially the whole feasible set of state-constrained optimal control problems. But, in addition, the authors are convinced that this methodology, if combined with sufficient optimality conditions, could turn out to be another conceptual approach to order of convergence proofs for numerical methods for the direct computation of optimal solutions; cf., e.g., [13, 12].

Naturally, convergence of the whole set of discrete solutions to the solution set of the continuous differential inclusion implies the convergence of the corresponding reachable sets. Hence, at least for set-valued Euler's method, we need not distinguish between these two aspects, but in this connection see the papers [24, 25], which extend the results in [11] for set-valued Euler's method to Runge–Kutta methods of order at least equal to 2 for problems without state constraints.

We denote by $\mathrm{AC}(I)$ the set of all absolutely continuous functions $y : I \to \mathbb{R}^n$ and by $\Theta : I \rightrightarrows \mathbb{R}^n$ a set-valued map with nonempty subsets of $\mathbb{R}^n$ as images.

PROBLEM 1.1. *Given an interval $I = [t_0, T]$, a nonempty set $Y_0 \subset \mathbb{R}^n$, and set-valued maps $F : I \times \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ and $\Theta : I \rightrightarrows \mathbb{R}^n$ with nonempty images, find all absolutely continuous solutions $y(\cdot)$ of the state-constrained differential inclusion (DIC)*

$$y'(t) \in F(t, y(t)) \qquad (\text{a.e. } t \in I), \tag{1.1}$$

$$y(t) \in \Theta(t) \qquad (t \in I), \tag{1.2}$$

$$y(t_0) = y_0 \in Y_0. \tag{1.3}$$

Clearly, we must have $y_0 \in \Theta(t_0)$ as well.

The unconstrained problem (DI) is given by (1.1), (1.3). *The set of solutions of (DI) and (DIC) is denoted by $\mathcal{Y}[T, t_0, Y_0]$, resp., $\mathcal{Y}^{\Theta}[T, t_0, Y_0]$.*

ALGORITHM 1.2. *Euler's method for (DIC) in Problem* 1.1 *with number of subintervals $N \in \mathbb{N}$ and step-size $h = \frac{T-t_0}{N}$ is given by*

$$\mathcal{Y}_N^{\Theta}[t_0, t_0, Y_0] := Y_0 \cap \Theta(t_0), \tag{1.4}$$

$$\mathcal{Y}_N^{\Theta}[t_{j+1}, t_0, Y_0] := \bigcup_{\eta_j \in \mathcal{Y}_N^{\Theta}[t_j, t_0, Y_0]} \big(\eta_j + hF(t_j, \eta_j)\big) \cap \Theta(t_{j+1}) \tag{1.5}$$

for $j = 0, \dots, N-1$.

Problem (DDIC) describes the solution of (1.4)–(1.5)*; its set of solutions is denoted by* $\mathcal{Y}_N^\Theta[T, t_0, Y_0]$. *In the absence of state constraints, the problem is called (DDI) and* $\mathcal{Y}_N[T, t_0, Y_0]$ *denotes the corresponding set of solutions.*

To measure distances, we define for $\eta = (\eta_j)_{j=0,\dots,N} \in \mathcal{Y}_N^\Theta[T, t_0, Y_0]$,

$$\operatorname{dist}_\infty(y(\cdot), \mathcal{Y}_N^\Theta[T, t_0, Y_0]) := \inf\left\{ \sup_{j=0,\dots,N} \|y(t_j) - \eta_j\| \,:\, \eta \in \mathcal{Y}_N^\Theta[T, t_0, Y_0] \right\},$$

$$\operatorname{dist}_\infty(\eta, \mathcal{Y}^\Theta[T, t_0, Y_0]) := \inf\left\{ \sup_{j=0,\dots,N} \|\eta_j - y(t_j)\| \,:\, y(\cdot) \in \mathcal{Y}^\Theta[T, t_0, Y_0] \right\},$$

$$\begin{aligned} \mathrm{d}_{H,\infty}(\mathcal{Y}^\Theta[T, t_0, Y_0], \mathcal{Y}_N^\Theta[T, t_0, Y_0]) := \max\Bigg\{ & \sup_{y(\cdot)\in\mathcal{Y}^\Theta[T,t_0,Y_0]} \operatorname{dist}_\infty(y(\cdot), \mathcal{Y}_N^\Theta[T, t_0, Y_0]), \\ & \sup_{\eta\in\mathcal{Y}_N^\Theta[T,t_0,Y_0]} \operatorname{dist}_\infty(\eta, \mathcal{Y}^\Theta[T, t_0, Y_0]) \Bigg\}. \end{aligned}$$

Here, the Euclidean vector norm on $\mathbb{R}^n$ is denoted by $\|\cdot\|$. For a subset $U \subset \mathbb{R}^n$, we denote by $\operatorname{dist}(x, U)$ the infimum of all Euclidean distances of the point $x \in \mathbb{R}^n$ to the points in U. We denote by $\mathrm{d}(U, V) = \sup_{u\in U} \operatorname{dist}(u, V)$ the one-sided Hausdorff distance from a subset $U \subset \mathbb{R}^n$ to another subset $V \subset \mathbb{R}^n$, and $\mathrm{d_H}(U, V)$ is the Hausdorff-distance defined as

$$\mathrm{d_H}(U, V) = \max\{\mathrm{d}(U, V), \mathrm{d}(V, U)\}.$$

We pose some of the following basic assumptions on the right-hand side:

(H1) F satisfies a linear growth condition, i.e., there exists $C \geq 0$ with

$$\|F(t, x)\| := \sup_{y\in F(t,x)} \|y\| \leq C(\|x\| + 1) \quad (t \in I,\, x \in \mathbb{R}^n).$$

(H2) F has nonempty, compact, convex images in $\mathbb{R}^n$.

(H3) F is Lipschitz in (t, x) for all $t \in I$, $x \in \mathbb{R}^n$ with constant $L \geq 0$, i.e.,

$$\mathrm{d_H}(F(s, x), F(t, y)) \leq L \cdot (|s - t| + \|x - y\|) \quad (s, t \in I,\, x, y \in \mathbb{R}^n).$$

The linear growth condition (H1) gives locally a boundedness of the images $F(t, x)$. A sufficient condition for (H1) is (H3) together with one bounded set $F(\widehat{t}, \widehat{x})$ (or (H2)). Condition (H2) is needed, since we want to apply the results from [11] for the unconstrained case. For practical applications, the Lipschitz condition could be restricted onto a compact set in which all values of all trajectories remain.

The following assumptions are required for the state constraints:

(C1) $\Theta : I \rightrightarrows \mathbb{R}^n$ has nonempty images explicitly given as

$$\Theta(t) := \{x \in \mathbb{R}^n \,:\, g(t, x) \leq 0\}$$

by a single scalar function $g : I \times \mathbb{R}^n \to \mathbb{R}$ which fulfills $g(\cdot, \cdot) \in \mathcal{C}^{1,L}(I \times \mathbb{R}^n)$, i.e., the derivative $\nabla g(\cdot, \cdot)$ is Lipschitz on $I \times \mathbb{R}^n$.
Furthermore, points $x \in \partial\Theta(t)$ with $t \in I$ are characterized by $g(t, x) = 0$.

(C2) The boundary of $\Theta(\cdot)$ fulfills the "strict inwardness condition" (cf. [15, 17, 18, 7]), i.e., there exists $\alpha, \mu > 0$ such that for all $(t, x) \in B_\mu(\operatorname{graph} \partial\Theta(\cdot)) \cap (I \times \mathbb{R}^n)$ it follows that

$$\min_{v\in F(t,x)} \left\langle \nabla g(t, x), \binom{1}{v} \right\rangle \leq -\alpha,$$

where

$$B_\mu(\operatorname{graph}\partial\Theta(\cdot)) = \left\{ \binom{t}{x} \in \mathbb{R}^{1+n} \,:\, \operatorname{dist}\left(\binom{t}{x}, \operatorname{graph}\partial\Theta(\cdot)\right) \le \mu \right\}.$$

From (C1) it follows that the images of $\Theta(\cdot)$ are closed. Existence of viable solutions could be proved under weaker assumptions; in this respect, cf. [16]. But since we are interested mainly in stability results, which require stronger assumptions anyway and imply existence as well, we will not discuss weaker existence results for the continuous and discrete case in this paper.

For the discrete situation in section 2, it is sufficient to pose weaker assumptions on $F(\cdot,\cdot)$ as follows:

(H1′) F satisfies a linear growth condition in integrable form, i.e., there exists a nonnegative function $C(\cdot) \in \mathcal{L}_1(I,\mathbb{R})$ with

$$\|F(t,x)\| := \sup_{y\in F(t,x)} \|y\| \le C(t)\cdot(\|x\|+1) \quad (t\in I,\, x\in\mathbb{R}^n).$$

(H2′) F has nonempty, closed images in $\mathbb{R}^n$.

(H3′) F is $L(t)$-Lipschitz in x for all $t\in I$ with $L(\cdot)\in\mathcal{L}_1(I,\mathbb{R})$, i.e.,

$$\mathrm{d_H}(F(t,x),F(t,y)) \le L(t)\cdot\|x-y\| \quad (x,y\in\mathbb{R}^n).$$

Usually, uniform boundedness of $C(\cdot)$ is assumed in (H1′), i.e., (H1). The same remark applies to $L(\cdot)$ in (H3′).

2. Stability for the unconstrained case. The essential stability result for differential inclusions without state constraints is given by the following theorem (for a complete proof, cf. [9, Lemma 8.3]).

THEOREM 2.1 (Gronwall–Filippov–Wazewski theorem). *Let $F(\cdot,\cdot)$ have closed images in $\mathbb{R}^n$, and let $Y_0\subset\mathbb{R}^n$ be nonempty and closed. For a given $\eta(\cdot)\in\mathrm{AC}(I)$ with*

$$\begin{aligned} \operatorname{dist}(\eta(t_0),Y_0) &\le \delta_0,\\ \operatorname{dist}(\eta'(t),F(t,\eta(t))) &\le \delta(t) \quad (\textit{a.e. } t\in I), \end{aligned}$$

with $\delta_0\ge 0$ and nonnegative $\delta(\cdot)\in\mathcal{L}_1(I,\mathbb{R})$, assume that

$$S := \{(t,x)\in I\times\mathbb{R}^n \,:\, \|x-\eta(t)\|\le\gamma\} \subset \operatorname{dom}(F)$$

for some $\gamma>\delta_0$. Let $F(\cdot,x)$ be measurable in t for all $x\in S$ and fulfill (H3′) *on S.*

Let $z(\cdot)$ be the solution of

$$\begin{aligned} z'(t) &= L(t)z(t)+\delta(t) \quad (\textit{a.e. } t\in I),\\ z(t_0) &= \delta_0. \end{aligned}$$

Then for all $\widetilde{T}\in I$ with $z(\widetilde{T})\le\gamma$ there exists a solution $y(\cdot)$ on $[t_0,\widetilde{T}]\subset I$ with

$$\begin{aligned} y'(t) &\in F(t,y(t)) \qquad (\textit{a.e. } t\in[t_0,\widetilde{T}]),\\ y(t_0) &= y_0\in Y_0, \end{aligned}$$

fulfilling the estimates

$$\begin{aligned} \|y(t)-\eta(t)\| &\le z(t) \qquad (t\in[t_0,\widetilde{T}]),\\ \|y'(t)-\eta'(t)\| &\le L(t)z(t)+\delta(t) \quad (\textit{a.e. } t\in[t_0,\widetilde{T}]), \end{aligned}$$

where

$$z(t) = e^{\int_{t_0}^{t} L(\sigma)\,d\sigma} \cdot \delta_0 + \int_{t_0}^{t} e^{\int_{\tau}^{t} L(\sigma)\,d\sigma} \cdot \delta(\tau)\,d\tau.$$

It will turn out in section 3 that Theorem 2.1, together with the following discrete analogue, is essential for the proof of stability for state-constrained differential inclusions.

THEOREM 2.2 (discrete Gronwall–Filippov–Wazewski theorem). *Let* $F : [t_0, T] \times \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ *fulfill* (H2′) *and* (H3′).

Consider the discrete difference inclusion

$$\frac{y_{k+1} - y_k}{h} \in F(t_k, y_k) \qquad (k = 0, \dots, N-1), \tag{2.1}$$

$$y_0 \in Y_0 \tag{2.2}$$

for a given $N \in \mathbb{N}$, *the step-size* $h = \frac{T - t_0}{N}$, *and a closed, nonempty starting set* $Y_0 \subset \mathbb{R}^n$.

Let $(\eta_k)_{k=0,\dots,N}$ *be a grid function with values in* $\mathbb{R}^n$ *and*

$$\begin{aligned} \operatorname{dist}(\eta_0, Y_0) &\le \delta_0, \\ \operatorname{dist}\left(\frac{\eta_{k+1} - \eta_k}{h}, F(t_k, \eta_k)\right) &\le \delta_{k+1} \qquad (k = 0, \dots, N-1). \end{aligned}$$

Abbreviate $L_k = L(t_k)$, $k = 0, \dots, N$, *and let* $(z_k)_{k=0,\dots,N} \subset \mathbb{R}$ *be the solution of*

$$\begin{aligned} \frac{z_{k+1} - z_k}{h} &= L_k z_k + \delta_{k+1} \qquad (k = 0, \dots, N-1), \\ z_0 &= \delta_0. \end{aligned} \tag{2.3}$$

Then there exists a solution $(y_k)_{k=0,\dots,N}$ *of the discrete problem* (2.1)–(2.2) *with*

$$\begin{aligned} \|\eta_k - y_k\| &\le z_k \qquad (k = 0, \dots, N), \\ \left\| \frac{\eta_{k+1} - \eta_k}{h} - \frac{y_{k+1} - y_k}{h} \right\| &\le L_k z_k + \delta_{k+1} \qquad (k = 0, \dots, N-1). \end{aligned}$$

Proof. Since $Y_0 \subset \mathbb{R}^n$ is nonempty, there exists $y \in Y_0$ with $\operatorname{dist}(\eta_0, Y_0) \le \|\eta_0 - y\| =: r$. Hence, the best approximation y_0 of η_0 in Y_0 coincides with that in the compact set $Y_0 \cap B_r(\eta_0)$, i.e.,

$$\|\eta_0 - y_0\| = \operatorname{dist}(\eta_0, Y_0) \le \delta_0 = z_0.$$

Assume that the assertion is true for $j = 0, \dots, k$, $k \in \{0, \dots, N-1\}$. Arguing as in the case $k = 0$, there exists $\xi_k^y \in F(t_k, y_k)$ for $\xi_k^\eta = \frac{1}{h}(\eta_{k+1} - \eta_k)$ with

$$\begin{aligned} &\|\xi_k^\eta - \xi_k^y\| = \operatorname{dist}(\xi_k^\eta, F(t_k, y_k)), \\ &\|\xi_k^\eta - \xi_k^y\| \le \operatorname{dist}(\xi_k^\eta, F(t_k, \eta_k)) + \mathrm{d_H}(F(t_k, \eta_k), F(t_k, y_k)) \le L_k \|\eta_k - y_k\| + \delta_{k+1}. \end{aligned}$$

Setting $y_{k+1} := y_k + h\xi_k^y$ yields

$$\begin{aligned} \|\eta_{k+1} - y_{k+1}\| &= \|(\eta_k + h\xi_k^\eta) - (y_k + h\xi_k^y)\| \le \|\eta_k - y_k\| + h\|\xi_k^\eta - \xi_k^y\| \\ &\le (1 + hL_k)\|\eta_k - y_k\| + h\delta_{k+1} \le (1 + hL_k)z_k + h\delta_{k+1} = z_{k+1}. \end{aligned}$$

∎

The explicit solution formula for the linear difference equation (2.3) yields immediately the following more specific estimates of the growth of the error bounds z_k $(k = 0, \dots, N)$.

COROLLARY 2.3. *With the assumptions as in Theorem* 2.2 *and for a Riemann integrable* $L(\cdot)$ *in* (H3′), *we can estimate the error bounds* z_k *for* $k = 0, \dots, N$ *as*

$$z_k = \delta_0 \cdot \prod_{\mu=0}^{k-1}(1 + hL_\mu) + h\sum_{j=1}^{k}\delta_j \cdot \prod_{\mu=j}^{k-1}(1 + hL_\mu),$$

$$\prod_{\mu=j}^{k-1}(1 + hL_\mu) \le \prod_{\mu=j}^{k-1} e^{hL_\mu} = e^{h\sum_{\mu=j}^{k-1} L_\mu} \le e^{C_L} \quad (j = 0, \dots, k), \tag{2.4}$$

where C_L *is an upper bound for the Riemann sums of the integral* $\int_{t_0}^{T} L(t)\,dt$.

If, furthermore, $L_k = L$ *for* $k = 0, \dots, N$, *then* $(1 + hL)^k \le e^{Lkh}$ *and for* $L > 0$,

$$z_k \le e^{Lkh}\delta_0 + \begin{cases} \frac{1}{L}(e^{Lkh} - 1) \cdot \max\limits_{j=1,\dots,k} \delta_j, \\ e^{L(k-1)h} \cdot h \sum\limits_{j=1}^{k} \delta_j. \end{cases} \tag{2.5}$$

The following lemmas are simple consequences of the growth condition and are well known in the literature (cf., e.g., [11, 19, 6]). They exhibit interesting connections between the continuous situation and the discrete situation in the case when $N \to \infty$.

LEMMA 2.4. *Let* $F(\cdot,\cdot)$ *satisfy* (H1′). *Then all solutions* $y(\cdot)$ *of (DI) in Problem* 1.1 *with bounded starting set* $Y_0 \subset \mathbb{R}^n$ *are uniformly bounded by* $M := (\|Y_0\| + C_L) \cdot (1 + C_L e^{C_L})$ *with* $C_L := \|C(\cdot)\|_{\mathcal{L}_1(I)}$ *and stay in a compactum* $S \subset \mathbb{R}^n$.

LEMMA 2.5. *Let* $F(\cdot,\cdot)$ *satisfy* (H1). *Then all solutions* $y(\cdot)$ *of (DI) in Problem* 1.1 *with bounded starting set* $Y_0 \subset \mathbb{R}^n$ *have a uniform Lipschitz constant.*

LEMMA 2.6. *Let* $F(\cdot,\cdot)$ *satisfy* (H1′) *with Riemann integrable* $C(\cdot)$, *and let* C_R *denote an upper bound for the Riemann sums. Then all solutions* $(\eta_k)_{k=0,\dots,N}$ *of (DDI) in Euler's method, Algorithm* 1.2, *with bounded starting set* $Y_0 \subset \mathbb{R}^n$ *are bounded uniformly in* $N \in \mathbb{N}$ *by* $M := (\|Y_0\| + C_R) \cdot (1 + C_R e^{C_R})$ *and stay in a compactum* $S \subset \mathbb{R}^n$.

Choosing $C_R = \|C(\cdot)\|_{\mathcal{L}_1(I)} + \varepsilon$ for all $N \ge N_0(\varepsilon)$ emphasizes the similarity of Lemma 2.6 to Lemma 2.4.

LEMMA 2.7. *Let* $F(\cdot,\cdot)$ *satisfy* (H1). *Then all solutions* $(\eta_k)_{k=0,\dots,N}$ *of (DDI) in Euler's method, Algorithm* 1.2, *with bounded starting set* $Y_0 \subset \mathbb{R}^n$ *have a Lipschitz constant uniformly in* $N \in \mathbb{N}$.

Proof. Let M be the bound for all discrete solutions $(\eta_k)_{k=0,\dots,N}$ according to Lemma 2.6. Then it follows for $N \in \mathbb{N}$ and $j, k \in \{0, 1, \dots, N\}$ with $j \le k$ that

$$\begin{aligned}\|\eta_k - \eta_j\| &= \left\|\sum_{\mu=j}^{k-1}(\eta_{\mu+1} - \eta_\mu)\right\| \le h\sum_{\mu=j}^{k-1}\left\|\frac{1}{h}(\eta_{\mu+1} - \eta_\mu)\right\| \le h\sum_{\mu=j}^{k-1}\|F(t_\mu, \eta_\mu)\| \\ &\le h\sum_{\mu=j}^{k-1} C(\|\eta_\mu\| + 1) \le C(M+1)(k-j)h = C(M+1)(t_k - t_j). \qquad \square\end{aligned}$$

3. Stability analysis for the state-constrained case. There are several variants in the literature of the Gronwall–Filippov–Wazewski theorem for the continuous

state-constrained case (cf. [15, Theorems 4.1 and 4.2], [17, Lemmas 3.3 and 4.4], [18, Theorem 3.1], as well as [7, Lemma 3.9], [8], and [23, Lemma 2.2(b)], which are all based on Soner's work in [22]). These variants were also denoted as theorems on the "existence of feasible neighboring trajectories" or as "tracking lemmas." Exemplarily, we treat here the so-called "smooth" case, where the function $g(t,x)$ determines the state constraints $\Theta(t)$ and $g(\cdot,\cdot) \in \mathcal{C}^{1,L}(I \times \mathbb{R}^n)$.

A typical result for the continuous situation is given in the following.

THEOREM 3.1. *Consider Problem* 1.1 *with time-dependent state constraint* $\Theta(\cdot)$. *Assume conditions* (H2)–(H3) *on the right-hand side* $F(\cdot,\cdot)$ *and conditions* (C1)–(C2) *on the state constraints.*

Then for every $y_0 \in \Theta(t_0)$ *there exists a positive constant* C *such that for every* $\eta(\cdot) \in \mathcal{Y}[T,t_0,y_0]$ *there exists* $y(\cdot) \in \mathcal{Y}^{\Theta}[T,t_0,y_0]$ *with*

$$\sup_{t\in[t_0,T]} \|\eta(t) - y(t)\| \le C \sup_{t\in[t_0,T]} \operatorname{dist}(\eta(t),\Theta(t)).$$

We will omit the proof of this theorem, since it exploits a similar strategy as in [15, Theorem 4.1], using in addition a result from [6, Theorem 3.2.4].

The reader should be aware that under considerably weaker assumptions, e.g., when no convexity is needed, Lipschitz condition with respect to both variables can be weakened, and analogous results for the continuous situation hold. However, the proof of the discrete analogue presented here could be given only under stronger assumptions until now. Contrary to the assumptions (HC_1)–(HC_4) in [6], we allow time-dependent state constraints even in the discrete situation and simplify the conditions for the error estimate.

In any case, we want to emphasize the fact that *both* stability results for the continuous and discrete case are needed for convergence of discrete approximations of state-constrained differential inclusions described in section 4.

We now present a rather detailed analysis of the discrete analogue of Theorem 3.1, partly following [6], but admitting time-dependent state constraints. We want to stress that this discrete analysis is in some respects rather technical but nevertheless essential for the convergence analysis in section 4. It would be very desirable to have available the discrete analogues of all those refined results of [15, Theorem 4.2], [17, Lemma 3.3], [18, Theorem 3.1] (smooth case), resp., [15, Theorem 4.1], [17, Lemma 4.4] (nonsmooth case), for the continuous situation. See also [17] for a detailed discussion of the smooth and nonsmooth cases.

THEOREM 3.2. *Consider problem (DDIC) in* (1.4)–(1.5) *with time-dependent state constraint* $\Theta(\cdot)$. *Assume conditions* (H2)–(H3) *on the right-hand side* $F(\cdot,\cdot)$ *and conditions* (C1)–(C2) *on the state constraints.*

Then for every $y_0 \in \Theta(t_0)$ *there exist* $N_0 \in \mathbb{N}$ *and a positive constant* C *such that for all* $N \ge N_0$ *and for all discrete solutions* $(\eta_k)_{k=0,\dots,N} \in \mathcal{Y}_N[T,t_0,y_0]$ *there exists a discrete solution* $(y_k)_{k=0,\dots,N} \in \mathcal{Y}_N^{\Theta}[T,t_0,y_0]$ *with*

$$\max_{k=0,\dots,N} \|\eta_k - y_k\| \le C \left(h + \max_{k=0,\dots,N} \operatorname{dist}(\eta_k,\Theta(t_k)) \right).$$

Proof. Consider an arbitrary, in general nonfeasible, solution $(\eta_k)_{k=0,\dots,N}$ and set

$$\delta_N := \max_{k=0,\dots,N} \operatorname{dist}(\eta_k,\Theta(t_k)).$$

Case A. Solution η_k is feasible for $k \in \mathcal{I} = \{0,\dots,N\}$.

Clearly, $\delta_N = 0$ and the assertion is valid for $y_k := \eta_k$, $k \in \mathcal{I}$.

Case B. Solution η_k is not feasible for some $k \in \mathcal{I}$.

In this case, $\delta_N > 0$. On a small index set $\mathcal{I}_0 = \{0, \dots, k_1\}$ with k_1 independent from $(\eta_k)_{k\in\mathcal{I}}$ the result will be proved as a first step.

Denote by L_η the uniform Lipschitz constant for all discrete solutions according to Lemma 2.7; by L, resp., $L_{\nabla g}$, the Lipschitz constant of $F(\cdot,\cdot)$, resp., $\nabla g(\cdot,\cdot)$; and choose the constants μ and α as in (C2). Without loss of generality, $L > 0$. Let M_2 be the maximum of $\|\nabla g(t,x)\|$ for $(t,x) \in I \times S$, with S being the compactum according to Lemma 2.6.

Define

$$\tau_1 := \max\left\{ t \in [t_0, T] \,:\, t \le t_0 + \frac{\mu}{2(L_\eta + 1)}, \right. \tag{3.1}$$

$$L_{\nabla g}(t - t_0) \le \frac{M_2}{2(L_\eta + 1)}, \tag{3.2}$$

$$\left. \max\left\{ M_2(L_\eta + 1), (L_\eta + 1)^2 \cdot \frac{L_{\nabla g}}{L} \right\} \cdot (e^{L(t-t_0)} - 1) \le \frac{\alpha}{12} \right\}, \tag{3.3}$$

which is independent of all discrete solutions and all $N \in \mathbb{N}$.[1]

For the discrete case, additional assumptions on the step-size are necessary to construct a viable solution.

Choose $N_0 \in \mathbb{N}$ with

$$h_{N_0} = \frac{T - t_0}{N_0} \le \tau_1 - t_0, \tag{3.4}$$

$$h_{N_0} \le \frac{\mu}{2(L_\eta + 1)}, \tag{3.5}$$

$$h_{N_0} L_{\nabla g} \le \frac{\alpha}{2(L_\eta + 1)^2}, \tag{3.6}$$

$$h_{N_0} L_{\nabla g} \le \frac{M_2}{L_\eta + 1}, \tag{3.7}$$

determining the maximal allowed step-size h_{N_0}.[2]

Inequality (3.4) is needed to guarantee that at least one step of Euler's method can be performed to reach a time not exceeding τ_1. Inequality (3.5) follows from (3.1) and (3.4). It ensures that a discrete solution, before violating the state constraints at the next index, will be sufficiently near the boundary such that there exists a direction which steers the solution into the interior. Inequalities (3.6)–(3.7) are needed to show the viability of the solution in this phase and control the error of Taylor expansions.

From now on, let $N \ge N_0$, $h = \frac{T-t_0}{N}$, and define in view of (3.4),

$$k_1 := \left\lfloor \frac{\tau_1 - t_0}{h} \right\rfloor \ge 1, \tag{3.8}$$

$$\hat{k}_1 := \min\{k \in \mathcal{I} \,:\, \eta_{k+1} \notin \Theta(t_{k+1})\} < N,$$

where k_1 is the biggest natural number not exceeding $\frac{\tau_1 - t_0}{h}$.

It is clear that $t_{k_1} \le \tau_1$ also satisfies the requirements in (3.1)–(3.3).

[1] Inequalities (3.1)–(3.3) are used in (3.14), (3.27), resp., in (3.25), (3.26).

[2] Inequalities (3.4)–(3.7) are used in (3.8), (3.10), (3.16), resp., (3.24).

Case B(i). $k_1 \le \hat{k}_1$, i.e., the solution η_k is feasible for $k \in \widetilde{\mathcal{I}}_0 := \{0, \dots, \hat{k}_1\} \supset \mathcal{I}_0$. Define

$$y_k := \eta_k \quad (k \in \mathcal{I}_0),$$

which fulfills the assertion on $\mathcal{I}_0$.

Case B(ii). $k_1 > \hat{k}_1$, i.e., the solution η_k is feasible for $k \in \widetilde{\mathcal{I}}_0 \subsetneqq \mathcal{I}_0$.

In the first phase, set

$$y_k := \eta_k \quad (k \in \widetilde{\mathcal{I}}_0). \tag{3.9}$$

Since $\eta_{\hat{k}_1} \in \partial\Theta(t_{\hat{k}_1})$ cannot be guaranteed in the discrete case (only $\eta_{\hat{k}_1} \in \Theta(t_{\hat{k}_1})$), the distance to the boundary must be estimated and should not exceed $\frac{\mu}{2}$ to guarantee an inward steering direction. The function $\varphi(s) = g(t_{\hat{k}_1} + s,\ \eta_{\hat{k}_1} + s\frac{\eta_{\hat{k}_1+1} - \eta_{\hat{k}_1}}{h})$ is continuous on $[0, h]$ with

$$\varphi(0) = g(t_{\hat{k}_1}, \eta_{\hat{k}_1}) \le 0, \quad \varphi(h) = g(t_{\hat{k}_1+1}, \eta_{\hat{k}_1+1}) > 0.$$

Therefore, there exists a zero $\bar{s} \in [0, h]$ of the function $\varphi(\cdot)$. Now, use (3.5) and (C1) to show

$$\begin{aligned} &\operatorname{dist}\left(\begin{pmatrix} t_{\hat{k}_1} \\ \eta_{\hat{k}_1} \end{pmatrix}, \operatorname{graph} \partial\Theta(\cdot)\right) \le \left\| \begin{pmatrix} t_{\hat{k}_1} \\ \eta_{\hat{k}_1} \end{pmatrix} - \begin{pmatrix} t_{\hat{k}_1} + \bar{s} \\ \eta_{\hat{k}_1} + \bar{s}\frac{\eta_{\hat{k}_1+1} - \eta_{\hat{k}_1}}{h} \end{pmatrix} \right\| \\ &\le \bar{s}\left(1 + \frac{1}{h} \cdot \|\eta_{\hat{k}_1+1} - \eta_{\hat{k}_1}\|\right) \le (1 + L_\eta)h \le \frac{\mu}{2}. \end{aligned} \tag{3.10}$$

Define (without loss of generality, the Lipschitz constant L_g of $g(\cdot)$ is greater than 0)

$$\kappa_1 := \min\left\{ \frac{k_1 - \hat{k}_1}{1 + \frac{\delta_N}{h}}, \frac{3}{\alpha}(L_g + 3M_2(L_\eta + 1)) \right\}, \tag{3.11}$$

$$\begin{aligned} \bar{\delta}_1 &:= \left\lfloor \kappa_1 \left(1 + \frac{\delta_N}{h}\right) + 1 \right\rfloor \ge 1, \\ \bar{k}_1 &:= \hat{k}_1 + \bar{\delta}_1, \end{aligned} \tag{3.12}$$

which determines the length of the inward steering phase $\widehat{\mathcal{I}}_0 := \{\hat{k}_1, \hat{k}_1 + 1, \dots, \bar{k}_1\} \subset \mathcal{I}_0$.[3] The nonnegative number κ_1 controls whether the corresponding time interval reaches t_{k_1} or guarantees the feasibility on the second time interval; $\bar{\delta}_1$ is the number of steps in the second phase in Case B(ii.1), resp., B(ii.2) following. Notice that κ_1 and $\bar{k}_1$ depend on the individual solution.

Consider the solution $(\widehat{y}_k)_{k \in \widehat{\mathcal{I}}_0}$ of the discrete inclusion

$$\begin{aligned} \frac{1}{h}(x_{k+1} - x_k) &\in Y(t_k, x_k) \quad (k \in \widehat{\mathcal{I}}_0 \setminus \{\bar{k}_1\}), \\ x_{\hat{k}_1} &= y_{\hat{k}_1} \end{aligned}$$

[3]The first term in (3.11) is used in (3.19), the second one in (3.29), while (3.12) is used in (3.28) and (3.33).

on the second index set $\widehat{\mathcal{I}}_0$. Here, $Y(t,x)$ is defined as follows:

$$\varphi(t,x) = \min_{v\in F(t,x)} \left\langle \nabla g(t,x), \binom{1}{v} \right\rangle, \tag{3.13}$$
$$Y(t,x) = \left\{ v \in F(t,x) \,:\, \left\langle \nabla g(x), \binom{1}{v} \right\rangle = \varphi(t,x) \right\},$$

where $\varphi(\cdot,\cdot)$ is continuous on $\operatorname{graph}\Theta(\cdot)$ by [3, Theorem 1.4.16] and $Y(t,x)$ has compact, nonempty images and is upper semicontinuous by [2, section 1.2, Theorem 6].

We choose $\hat{k}_1$ so that inward steering is possible and show that this is the case for all $k \in \widehat{\mathcal{I}}_0$ as well. From the Lipschitz continuity of all discrete solutions by Lemma 2.7 and (3.10), we get for $k \in \widehat{\mathcal{I}}_0$ that

$$\begin{aligned} \|\widehat{y}_k - \widehat{y}_{\hat{k}_1}\| &\le L_\eta (k-\hat{k}_1)h, \\ \operatorname{dist}\left(\binom{t_k}{\widehat{y}_k}, \operatorname{graph}\partial\Theta(\cdot)\right) &\le \left\| \binom{t_k}{\widehat{y}_k} - \binom{t_{\hat{k}_1}}{\widehat{y}_{\hat{k}_1}} \right\| + \operatorname{dist}\left(\binom{t_{\hat{k}_1}}{\widehat{y}_{\hat{k}_1}}, \operatorname{graph}\partial\Theta(\cdot)\right) \\ &\le |t_k - t_{\hat{k}_1}| + \|\widehat{y}_k - \widehat{y}_{\hat{k}_1}\| + \frac{\mu}{2}. \end{aligned}$$

Estimate $(k-\hat{k}_1)h$ by $t_{k_1} - t_0$ and use (3.1) to show

$$\operatorname{dist}\left(\binom{t_k}{\widehat{y}_k}, \operatorname{graph}\partial\Theta(\cdot)\right) \le (L_\eta + 1)(k-\hat{k}_1)h + \frac{\mu}{2} \le \mu. \tag{3.14}$$

The proof of the feasibility of $(\widehat{y}_k)_{k\in\widehat{\mathcal{I}}_0}$ is not as simple as in the continuous case. Since $\widehat{y}_{\hat{k}_1} \in \Theta(t_{\hat{k}_1})$ per definition, we have $g(t_{\hat{k}_1}, \widehat{y}_{\hat{k}_1}) \le 0$ and

$$g(t_k, \widehat{y}_k) \le g(t_k, \widehat{y}_k) - g(t_{\hat{k}_1}, \widehat{y}_{\hat{k}_1}) = \sum_{j=\hat{k}_1}^{k-1} (g(t_{j+1}, \widehat{y}_{j+1}) - g(t_j, \widehat{y}_j)).$$

Set $\psi(s) = g(t_j + sh,\ \widehat{y}_j + s(\widehat{y}_{j+1} - \widehat{y}_j))$ for $s \in [0,1]$ and some $j \in \widehat{\mathcal{I}}_0$; then a Taylor expansion up to terms of order 1 yields, by the Lipschitz continuity of $\nabla g(\cdot,\cdot)$,

$$g(t_{j+1}, \widehat{y}_{j+1}) \le g(t_j, \widehat{y}_j) + \left\langle \nabla g(t_j, \widehat{y}_j), \binom{h}{\widehat{y}_{j+1} - \widehat{y}_j} \right\rangle + L_{\nabla g}(L_\eta + 1)^2 h^2. \tag{3.15}$$

Hence, due to (3.6) it follows that

$$\begin{aligned} g(t_k, \widehat{y}_k) &\le \sum_{j=\hat{k}_1}^{k-1} h \left\langle \nabla g(t_j, \widehat{y}_j), \binom{1}{\frac{\widehat{y}_{j+1} - \widehat{y}_j}{h}} \right\rangle + (L_{\nabla g}(L_\eta + 1)^2 h) \cdot (k - \hat{k}_1)h \\ &\le \sum_{j=\hat{k}_1}^{k-1} h \left\langle \nabla g(t_j, \widehat{y}_j), \binom{1}{\frac{\widehat{y}_{j+1} - \widehat{y}_j}{h}} \right\rangle + \frac{\alpha}{2} \cdot (k - \hat{k}_1)h. \end{aligned} \tag{3.16}$$

Using (C2) due to (3.14) and $\frac{\widehat{y}_{j+1} - \widehat{y}_j}{h} \in Y(t_j, \widehat{y}_j)$ together with (3.13), we progress to the inequalities

$$g(t_k, \widehat{y}_k) \le h \sum_{j=\hat{k}_1}^{k-1} \varphi(t_j, \widehat{y}_j) + \frac{\alpha}{2} \cdot (k - \hat{k}_1)h \le -\frac{\alpha}{2} \cdot (k - \hat{k}_1)h. \tag{3.17}$$

Therefore, we have finally proved that $\widehat{y}_k \in \Theta(t_k)$ and

$$\|\widehat{y}_k - \eta_k\| \le \|\widehat{y}_k - y_{\hat{k}_1}\| + \|\eta_{\hat{k}_1} - \eta_k\| \le 2L_\eta(k-\hat{k}_1)h \le 2L_\eta\bar{\delta}_1 h \quad (k \in \widehat{\mathcal{I}}_0). \tag{3.18}$$

Case B(ii.1). The inward steering phase reaches the end of index set $\mathcal{I}_0$.

If $\bar{k}_1 = \hat{k}_1 + \bar{\delta}_1 = k_1$, then the definition of the constructed solution is continued to $\widehat{\mathcal{I}}_0$ as

$$y_k := \widehat{y}_k \quad (k \in \widehat{\mathcal{I}}_0 \setminus \{\hat{k}_1\}),$$

so that the claim is verified on $\widehat{\mathcal{I}}_0$ and therefore also on $\mathcal{I}_0$.

Case B(ii.2). The Filippov solution follows the time-delayed solution for the rest of the indices in $\mathcal{I}_0 \setminus \widehat{\mathcal{I}}_0$.

Now $\bar{k}_1 = \hat{k}_1 + \bar{\delta}_1 < k_1$; set $\bar{\mathcal{I}}_0 := \{\bar{k}_1, \bar{k}_1 + 1, \dots, k_1\}$. From $\kappa_1(1 + \frac{\delta_N}{h}) < \bar{\delta}_1$ it follows that $\kappa_1 = \frac{3}{\alpha}(L_g + 3M_2(L_\eta + 1))$ since

$$\kappa_1 < \frac{k_1 - \hat{k}_1}{1 + \frac{\delta_N}{h}}. \tag{3.19}$$

Consider the Filippov solution $(\bar{y}_k)_{k\in\bar{\mathcal{I}}_0}$ of

$$\frac{1}{h}(x_{k+1} - x_k) \in F(t_k, x_k) \quad (k \in \bar{\mathcal{I}}_0 \setminus \{k_1\}),$$
$$x_{\bar{k}_1} = y_{\bar{k}_1}$$

following the solution $(\eta_{k-\bar{\delta}_1})_{k\in\bar{\mathcal{I}}_0}$. Since the discrete version of Filippov's theorem, Theorem 2.2, will be applied, we study the following error terms:

$$\begin{aligned} &\|\bar{y}_{\bar{k}_1} - \eta_{\bar{k}_1-\bar{\delta}_1}\| = \|y_{\bar{k}_1} - \eta_{\hat{k}_1}\| = \|y_{\bar{k}_1} - y_{\hat{k}_1}\| \le L_\eta\bar{\delta}_1 h, \\ &\operatorname{dist}\Bigg(\underbrace{\frac{1}{h}(\eta_{k+1-\bar{\delta}_1} - \eta_{k-\bar{\delta}_1})}_{\in F(t_{k-\bar{\delta}_1}, \eta_{k-\bar{\delta}_1})}, F(t_k, \eta_{k-\bar{\delta}_1})\Bigg) \le L\bar{\delta}_1 h. \end{aligned} \tag{3.20}$$

The time delay $\bar{\delta}_1$ not only helps in (3.20), since $\eta_{\bar{k}_1-\bar{\delta}_1}$ coincides with $y_{\hat{k}_1}$, but also allows us to reuse the estimates on the second index set $\widehat{\mathcal{I}}_0$ (namely (3.18)) for the starting values on the third index set. For the distance to the right-hand side of the difference inclusion, the Lipschitz continuity of $F(\cdot,\cdot)$ with respect to t was used. The discrete Filippov theorem, Theorem 2.2, together with Corollary 2.3, finally establishes the estimates

$$\begin{aligned} \|\bar{y}_k - \eta_{k-\bar{\delta}_1}\| &\le (1+hL)^{k-\bar{k}_1} L_\eta\bar{\delta}_1 h + ((1+hL)^{k-\bar{k}_1} - 1)\bar{\delta}_1 h \\ &= ((L_\eta + 1)(1+hL)^{k-\bar{k}_1} - 1)\bar{\delta}_1 h, \end{aligned} \tag{3.21}$$

$$\begin{aligned} &\left\|\frac{1}{h}(\eta_{k+1-\bar{\delta}_1} - \eta_{k-\bar{\delta}_1}) - \frac{1}{h}(\bar{y}_{k+1} - \bar{y}_k)\right\| \\ &\qquad \le L(L_\eta + 1)(1+hL)^{k-\bar{k}_1}\bar{\delta}_1 h \end{aligned} \tag{3.22}$$

on $\bar{I}_0$. They are used twice: first to estimate the deviation of the feasible solution to the given one in

$$\begin{aligned} \|\bar{y}_k - \eta_k\| &\le \|\bar{y}_k - \eta_{k-\bar{\delta}_1}\| + \|\eta_{k-\bar{\delta}_1} - \eta_k\| \\ &\le \Big((L_\eta + 1)e^{L(k-\bar{k}_1)h} + L_\eta - 1\Big)\bar{\delta}_1 h \end{aligned} \tag{3.23}$$

and second to show feasibility. To this purpose, the state constraint is split into four terms for each $k \in \bar{\mathcal{I}}_0$. Hereby, the Taylor expansion as in (3.15) will be used:

$$\begin{aligned}
g(t_k,\bar{y}_k) &= \underbrace{g(t_{\bar{k}_1},\bar{y}_{\bar{k}_1})}_{=T_A} + \underbrace{g(t_{k-\bar{\delta}_1},\eta_{k-\bar{\delta}_1}) - g(t_{\bar{k}_1-\bar{\delta}_1},\eta_{\bar{k}_1-\bar{\delta}_1})}_{=T_B} \\
&\quad + \sum_{j=\bar{k}_1}^{k-1} \big(g(t_{j+1},\bar{y}_{j+1}) - g(t_j,\bar{y}_j)\big) \\
&\quad - \sum_{j=\bar{k}_1}^{k-1} \big(g(t_{j+1-\bar{\delta}_1},\eta_{j+1-\bar{\delta}_1}) - g(t_{j-\bar{\delta}_1},\eta_{j-\bar{\delta}_1})\big) \\
&\le T_A + T_B + h \sum_{j=\bar{k}_1}^{k-1} \left\langle \nabla g(t_j,\bar{y}_j), \begin{pmatrix} 1 \\ \frac{\bar{y}_{j+1}-\bar{y}_j}{h} \end{pmatrix} \right\rangle + L_{\nabla g}(L_\eta+1)^2(k-\bar{k}_1)h^2 \\
&\quad - h \sum_{j=\bar{k}_1}^{k-1} \left\langle \nabla g(t_{j-\bar{\delta}_1},\eta_{j-\bar{\delta}_1}), \begin{pmatrix} 1 \\ \frac{\eta_{j+1-\bar{\delta}_1}-\eta_{j-\bar{\delta}_1}}{h} \end{pmatrix} \right\rangle + L_{\nabla g}(L_\eta+1)^2(k-\bar{k}_1)h^2 \\
&= T_A + T_B + \underbrace{h \sum_{j=\bar{k}_1}^{k-1} \left\langle \nabla g(t_j,\bar{y}_j), \begin{pmatrix} 1 \\ \frac{\bar{y}_{j+1}-\bar{y}_j}{h} \end{pmatrix} - \begin{pmatrix} 1 \\ \frac{\eta_{j+1-\bar{\delta}_1}-\eta_{j-\bar{\delta}_1}}{h} \end{pmatrix} \right\rangle}_{=T_C} \\
&\quad + \underbrace{h \sum_{j=\bar{k}_1}^{k-1} \left\langle \nabla g(t_j,\bar{y}_j) - \nabla g(t_{j-\bar{\delta}_1},\eta_{j-\bar{\delta}_1}), \begin{pmatrix} 1 \\ \frac{\eta_{j+1-\bar{\delta}_1}-\eta_{j-\bar{\delta}_1}}{h} \end{pmatrix} \right\rangle}_{=T_D} \\
&\quad + \underbrace{2L_{\nabla g}(L_\eta+1)^2(k-\bar{k}_1)h^2}_{=T_E} = T_A + T_B + T_C + T_D + T_E.
\end{aligned}$$

The next task will be to estimate each term separately. We estimate

$$T_A = g(t_{\bar{k}_1},\widehat{y}_{\bar{k}_1}) \le -\frac{\alpha}{2}\bar{\delta}_1 h$$

by (3.17), the corresponding inequality on the second index set.

The treatment of the second term is slightly more complicated, as in the continuous case, since we cannot assume that $g(t_{\hat{k}_1},\eta_{\hat{k}_1}) = 0$. Nevertheless, we know that at index $\hat{k}_1$ we are close to the boundary and at the next index $\hat{k}_1+1$ the iterate violates the state constraints so that

$$T_B = g(t_{k-\bar{\delta}_1},\eta_{k-\bar{\delta}_1}) - g(t_{\hat{k}_1},\eta_{\hat{k}_1}) < g(t_{k-\bar{\delta}_1},\eta_{k-\bar{\delta}_1}) + \underbrace{g(t_{\hat{k}_1+1},\eta_{\hat{k}_1+1})}_{>0} - g(t_{\hat{k}_1},\eta_{\hat{k}_1}).$$

The difference of the last two terms could be estimated as in (3.15):

$$\begin{aligned}
g(t_{\hat{k}_1+1},\eta_{\hat{k}_1+1}) - g(t_{\hat{k}_1},\eta_{\hat{k}_1}) &\le h\|\nabla g(t_{\hat{k}_1},\eta_{\hat{k}_1})\| \cdot \left(1 + \left\|\frac{\eta_{\hat{k}_1+1}-\eta_{\hat{k}_1}}{h}\right\|\right) \\
&\quad + L_{\nabla g}(L_\eta+1)^2h^2 \le \underbrace{\max_{(t,x)\in I\times S}\|\nabla g(t,x)\|}_{=M_2} \cdot (1+L_\eta)h + L_{\nabla g}(L_\eta+1)^2h^2,
\end{aligned}$$

where we used again the fact that all discrete solutions are contained within a compactum S by Lemma 2.6 and that all discrete solutions have a uniform Lipschitz constant L_η by Lemma 2.7. Mimicking the proof in the continuous case, we distinguish two cases to treat the first term in T_B.

If $\eta_{k-\bar\delta_1} \in \Theta(t_{k-\bar\delta_1})$, then $g(t_{k-\bar\delta_1}, \eta_{k-\bar\delta_1}) \le 0$ so that this first term has an advantageous sign. Otherwise, we introduce the projection $\eta^\pi_{k-\bar\delta_1} \in \partial\Theta(t_{k-\bar\delta_1})$ and estimate by using the definition of δ_N:

$$\begin{aligned}|g(t_{k-\bar\delta_1}, \eta_{k-\bar\delta_1}) - g(t_{k-\bar\delta_1}, \eta^\pi_{k-\bar\delta_1})| &\le L_g \|\eta_{k-\bar\delta_1} - \eta^\pi_{k-\bar\delta_1}\| \\ &= L_g \operatorname{dist}(\eta_{k-\bar\delta_1}, \Theta(t_{k-\bar\delta_1})) \le L_g\delta_N.\end{aligned}$$

In both cases, due to (3.7) we have

$$T_B \le L_g\delta_N + M_2 \cdot (1 + L_\eta)h + L_{\nabla g}(L_\eta + 1)^2h^2 \le L_g\delta_N + 2M_2 \cdot (1 + L_\eta)h. \tag{3.24}$$

In term T_C, the difference quotient of both solutions is compared, which was estimated in (3.22) by the discrete Filippov theorem. Moreover, the boundedness of the discrete solutions and the continuity of $\nabla g(\cdot,\cdot)$ are used, yielding

$$\begin{aligned}T_C &\le h \sum_{j=\bar k_1}^{k-1} \|\nabla g(t_j, \bar y_j)\| \cdot \left\| \frac{\bar y_{j+1} - \bar y_j}{h} - \frac{\eta_{j+1-\bar\delta_1} - \eta_{j-\bar\delta_1}}{h} \right\| \\ &\le M_2 h \sum_{j=\bar k_1}^{k-1} \left(L(L_\eta + 1)(1 + hL)^{j-\bar k_1}\bar\delta_1 h \right) = M_2(L_\eta + 1)((1 + hL)^{k-\bar k_1} - 1)\bar\delta_1 h.\end{aligned}$$

Since $(1 + hL)^{k-\bar k_1}$ can be estimated by Corollary 2.3 as $e^{L(k-\bar k_1)h} \le e^{Lk_1h} \le e^{L(\tau_1-t_0)}$, we can exploit that τ_1 was suitably chosen by (3.3), and we get

$$T_C \le \frac{\alpha}{12}\bar\delta_1 h. \tag{3.25}$$

The same estimate will be reached for the term T_D. The main keys are the Lipschitz continuity of $\nabla g(\cdot,\cdot)$, the uniform Lipschitz constant for all discrete solutions, and the estimates (3.21) from the discrete Filippov theorem, together with the estimate in (2.5):

$$\begin{aligned}T_D &\le h \sum_{j=\bar k_1}^{k-1} \|\nabla g(t_j, \bar y_j) - \nabla g(t_{j-\bar\delta_1}, \eta_{j-\bar\delta_1})\| \cdot \left(1 + \left\| \frac{\eta_{j+1-\bar\delta_1} - \eta_{j-\bar\delta_1}}{h} \right\| \right) \\ &\le h \sum_{j=\bar k_1}^{k-1} L_{\nabla g}(|t_j - t_{j-\bar\delta_1}| + \|\bar y_j - \eta_{j-\bar\delta_1}\|) \cdot (1 + L_\eta) \\ &\le (L_\eta + 1)L_{\nabla g} h \sum_{j=\bar k_1}^{k-1} (1 + (L_\eta + 1)(1 + hL)^{j-\bar k_1} - 1) \cdot \bar\delta_1 h \\ &\le (L_\eta + 1)L_{\nabla g} \frac{L_\eta + 1}{L} hL \sum_{j=\bar k_1}^{k-1} (1 + hL)^{j-\bar k_1} \cdot \bar\delta_1 h \\ &\le (L_\eta + 1)^2 \frac{L_{\nabla g}}{L}((1 + hL)^{k-\bar k_1} - 1) \cdot \bar\delta_1 h.\end{aligned}$$

Now, the reasoning is the same as for the term T_C, and hence

$$T_D \le \frac{\alpha}{12}\bar{\delta}_1 h. \tag{3.26}$$

For the estimation of T_E we need (3.2):

$$\begin{aligned} T_E &= 2L_{\nabla g}(L_\eta+1)^2(k-\hat{k}_1)h^2 \le 2L_{\nabla g}(L_\eta+1)^2(t_k - t_{\hat{k}_1})h \\ &\le 2L_{\nabla g}(L_\eta+1)^2(\tau_1 - t_0)h \le M_2(L_\eta+1)h. \end{aligned} \tag{3.27}$$

Now, we put all estimates together to show the feasibility. We have

$$\begin{aligned} g(t_k,\bar{y}_k) &\le T_A + T_C + T_D + T_B + T_E \le -\frac{\alpha}{2}\bar{\delta}_1 h + 2\cdot\frac{\alpha}{12}\bar{\delta}_1 h + T_B + T_E \\ &\le -\frac{\alpha}{3}\bar{\delta}_1 h + T_B + T_E. \end{aligned}$$

The definition (3.12) for $\bar{\delta}_1$ and $\kappa_1 = \frac{3}{\alpha}(L_g + 3M_2(L_\eta+1))$ yield

$$\frac{\alpha}{3}\bar{\delta}_1 h \ge \frac{\alpha}{3}\kappa_1\left(1+\frac{\delta_N}{h}\right)h \tag{3.28}$$

$$= (L_g + 3M_2(L_\eta+1))(h+\delta_N) \ge L_g\delta_N + 3M_2(L_\eta+1)h, \tag{3.29}$$

and hence the problematic term $L_g\delta_N$ could be eliminated by

$$\begin{aligned} g(t_k,\bar{y}_k) &\le -L_g\delta_N - 3M_2(L_\eta+1)h + L_g\delta_N + 2M_2(L_\eta+1)h \\ &\quad + M_2(L_\eta+1)h \le 0. \end{aligned} \tag{3.30}$$

Extend the feasible solution in the third phase to $\mathcal{I}_0$ by

$$y_k := \bar{y}_k \quad (k \in \bar{\mathcal{I}}_0 \setminus \{\bar{k}_1\}). \tag{3.31}$$

For all $k \in \mathcal{I}_0$, (3.9) and the estimates (3.18), (3.23) yield altogether

$$\|y_k - \eta_k\| \le \max\{2L_\eta, \underbrace{(L_\eta+1)e^{L(\tau_1-t_0)} + L_\eta - 1}_{=:M_3 \ge 2L_\eta}\}\cdot\bar{\delta}_1 h. \tag{3.32}$$

In the last inequality, $(k_1 - \bar{k}_1)h$ was estimated by $k_1 h \le \tau_1 - t_0$. Moreover,

$$\begin{aligned} \bar{\delta}_1 h &= \left\lfloor \kappa_1\left(1+\frac{\delta_N}{h}\right)+1\right\rfloor\cdot h \le \left(\kappa_1\left(1+\frac{\delta_N}{h}\right)+1\right)h \\ &\le \left(\frac{3}{\alpha}(L_g + 3M_2(L_\eta+1))\right)(h+\delta_N) + h = \mathcal{O}(h+\delta_N), \end{aligned} \tag{3.33}$$

$$\|y_k - \eta_k\| \le M_3\bar{\delta}_1 h \le \underbrace{M_3\left(1+\frac{3}{\alpha}(L_g+3M_2(L_\eta+1))\right)}_{=:\widetilde{M}}(h+\delta_N) = \mathcal{O}(h+\delta_N).$$

Extension to the whole index set $\mathcal{I}$. This process is well explained in the proof of [6, Theorem 3.2.6]: Divide the index set into J subsets with k_1 elements and set $\mathcal{I}_j := \{k_j, k_j+1, \dots, k_{j+1}\} \cap \{0,\dots,N\}$ with $k_j = jk_1$, $j = 0,\dots,J$.

(i) *First index set.* For $j=0$ the solution y_k is already constructed for $\mathcal{I}_0$. Set $\widetilde{C}_0 := 1 + \frac{\delta_N}{h}$ and $\Delta_0 = \lfloor \kappa_1\widetilde{C}_0 + 1\rfloor$.

(ii) *Recursive approach.* For $j > 0$ start the process by taking the end value of the feasible solution $y_{j\cdot k_1}$ on $\mathcal{I}_{j-1}$ as the starting value for the next iteration. Now, apply again the discrete Filippov theorem to construct the (in general, nonfeasible) solution $(z_k^{(j)})_{k\in\mathcal{I}_j}$ of

$$\frac{1}{h}(x_{k+1} - x_k) \in F(t_k, x_k) \quad (k \in \mathcal{I}_j),$$
$$x_{k_j} = y_{k_j}$$

that follows the nonfeasible solution $(\eta_k)_{k\in\mathcal{I}_j}$. The error term is governed by the difference of the starting values. Now, construct a feasible solution $(y_k)_{k\in\mathcal{I}_j}$ from $(z_k^{(j)})_{k\in\mathcal{I}_j}$. Then show that the deviation from $(y_k)_{k\in\mathcal{I}_j}$ to $(\eta_k)_{k\in\mathcal{I}_j}$ could be estimated by

$$\|y_k - \eta_k\| \le \widetilde{M} \sum_{\nu=0}^{j} e^{(j-\nu)Lk_1h} \Delta_\nu h \quad (k \in \mathcal{I}_j),$$

where for $j = 1, \dots, J$,

$$\widetilde{C}_j = \widetilde{C}_0 + \widetilde{M} \sum_{\nu=0}^{j-1} e^{(j-\nu)Lk_1h}, \quad \Delta_j = \lfloor \kappa_1 \widetilde{C}_j + 1 \rfloor.$$

Estimate J uniformly for all $N \in \mathbb{N}$ by $\lfloor \frac{T-t_0}{\tau_1 - h_{N_0}} + 1 \rfloor$ so that we finally prove the overall order $\mathcal{O}(h + \delta_N)$. ∎

Remark 3.3. Assume that $\Theta : I \Rightarrow \mathbb{R}^n$ with images in $\mathcal{C}(\mathbb{R}^n)$ has a $\mathcal{C}^{1,L}$-signed distance function

$$\widetilde{d}(t,x) := \begin{cases} \operatorname{dist}(x, \partial\Theta(t)) & \text{if } x \in \Theta(t), \\ -\operatorname{dist}(x, \partial\Theta(t)) = -\operatorname{dist}(x, \Theta(t)) & \text{if } x \in \mathbb{R}^n \setminus \Theta(t). \end{cases}$$

Then $\Theta(t) = \{x \in \mathbb{R}^n : -\widetilde{d}(t,x) \le 0\}$ fulfills the assumptions of Theorem 3.2.

4. Convergence analysis. Combining the stability results from section 3 for the continuous and discrete case, we are now in a position to prove order of convergence results for the discrete approximation of the set of all viable solutions of the differential inclusion by all viable discrete solutions.

An essential tool is the following result for differential inclusions without state constraints (cf. [11, section 1, Theorem]) which we formulate under stronger assumptions that will be needed later. The convexity is an important assumption for the convergence of Euler's method.

PROPOSITION 4.1. *Choose a compactum $S \subset \mathbb{R}^n$ containing all solutions of* (1.1), (1.3). *Let $F(\cdot,\cdot)$ fulfill* (H2)–(H3) *on S and let $Y_0 = \{y_0\}$.*

Then there exists a positive constant C such that for all $N \in \mathbb{N}$,

$$\mathrm{d}_{H,\infty}(\mathcal{Y}[T, t_0, y_0], \mathcal{Y}_N[T, t_0, y_0]) \le Ch.$$

The stability results from section 3 (Theorem 3.1 for the continuous case and Theorem 3.2 for the discrete case) are essential for the convergence proof of Euler's discretization of differential inclusions with state constraints.

THEOREM 4.2. *Assume hypotheses* (H2)–(H3) *together with* (C1)–(C2) *and let* $Y_0 = \{y_0\}$ *with* $y_0 \in \Theta(t_0)$.

Then there exist a positive constant C *and* $N_0 \in \mathbb{N}$ *such that for all* $N \geq N_0$,

$$\mathrm{d}_{H,\infty}(\mathcal{Y}^{\Theta}[T,t_0,y_0], \mathcal{Y}_N^{\Theta}[T,t_0,y_0]) \leq Ch.$$

Proof. This proof will use the notation of some constants from the proof of Theorem 3.2. Choose $N_0 \in \mathbb{N}$ from this theorem and $N \geq N_0$ so that additionally $h_{N_0} \leq \mu$ and $(C(M+1)+1)^2 L_{\nabla g} h_{N_0} \leq \frac{\alpha}{2}$, where M is the bound in Lemma 2.6 and α, μ are as in (C2).

Let us first construct a close discrete solution to a given $y(.) \in \mathcal{Y}^{\Theta}[T,t_0,y_0]$ to estimate the one-sided distance. According to Proposition 4.1, there exists $(\widetilde{\eta}_k)_{k=0,\dots,N} \in \mathcal{Y}_N[T,t_0,y_0]$ with

$$\max_{k=0,\dots,N} \|y(t_k) - \widetilde{\eta}_k\| \leq \widetilde{C}_1 h.$$

Since

$$\operatorname{dist}(\widetilde{\eta}_k, \Theta(t_k)) \leq \|\widetilde{\eta}_k - y(t_k)\| + \operatorname{dist}(y(t_k), \Theta(t_k)) \leq \widetilde{C}_1 h,$$

a solution $(\eta_k)_{k=0,\dots,N} \in \mathcal{Y}_N^{\Theta}[T,t_0,y_0]$ can be constructed by Theorem 3.2 with

$$\max_{k=0,\dots,N} \|\eta_k - \widetilde{\eta}_k\| \leq \widetilde{C}_2 h.$$

Hence, the grid function $y_N := (y(t_k))_{k=0,\dots,N}$ fulfills

$$\|\eta_k - y(t_k)\| \leq \|\eta_k - \widetilde{\eta}_k\| + \|\widetilde{\eta}_k - y(t_k)\| \leq (\widetilde{C}_1 + \widetilde{C}_2)h,$$
$$\operatorname{dist}_{\infty}(y_N, \mathcal{Y}_N^{\Theta}[T,t_0,y_0]) \leq (\widetilde{C}_1 + \widetilde{C}_2)h.$$

On the other hand, for a given discrete solution $\eta := (\eta_k)_{k=0,\dots,N} \in \mathcal{Y}_N^{\Theta}[T,t_0,y_0]$ one has to estimate the other one-sided distance. Proposition 4.1 shows the existence of $\widetilde{y}(\cdot) \in \mathcal{Y}[T,t_0,y_0]$ with

$$\max_{k=0,\dots,N} \|\eta_k - \widetilde{y}(t_k)\| \leq \widetilde{C}_1 h.$$

The reasoning is now more complicated since we need to estimate the following distance for all $t \in [t_k, t_{k+1}]$ and all $k \in \{0,\dots,N-1\}$:

$$\operatorname{dist}(\widetilde{y}(t), \Theta(t)) \leq \|\widetilde{y}(t) - \widetilde{y}(t_k)\| + \|\widetilde{y}(t_k) - \eta_k\| + \operatorname{dist}(\eta_k, \Theta(t)). \tag{4.1}$$

Since $\eta_k \in \Theta(t_k)$, the inequality $g(t_k, \eta_k) \leq 0$ holds.

(i) If $\binom{t_k}{\eta_k} \in B_\mu(\operatorname{graph} \partial\Theta(\cdot))$, then there exists $v_k \in F(t_k, \eta_k)$ by (C2) with

$$\left\langle \nabla g(t_k, \eta_k), \binom{1}{v_k} \right\rangle \leq -\alpha.$$

For $t \in [t_k, t_{k+1}]$, we set $\eta(t) := \eta_k + (t - t_k)v_k$ and consider

$$\begin{aligned} g(t,\eta(t)) &= g(t_k,\eta_k) + \int_{t_k}^{t} \frac{\mathrm{d}}{\mathrm{ds}} g(s,\eta(s))\, ds \leq \int_{t_k}^{t} \left\langle \nabla g(s,\eta(s)), \binom{1}{v_k} \right\rangle ds \\ &= \int_{t_k}^{t} \left\langle \nabla g(t_k,\eta_k), \binom{1}{v_k} \right\rangle ds + \int_{t_k}^{t} \left\langle \nabla g(s,\eta(s)) - \nabla g(t_k,\eta_k), \binom{1}{v_k} \right\rangle ds \\ &\leq -\alpha(t-t_k) + \int_{t_k}^{t} \|\nabla g(s,\eta(s)) - \nabla g(t_k,\eta_k)\| \cdot (1 + \|v_k\|)\, ds. \end{aligned}$$

Let us estimate both terms using (H1) and Lemma 2.6 by

$$\begin{aligned} 1 + \|v_k\| &\le 1 + \|F(t_k, \eta_k)\| \le 1 + C(\|\eta_k\| + 1) \le C(M+1) + 1, \\ \|\nabla g(s, \eta(s)) - \nabla g(t_k, \eta_k)\| &\le L_{\nabla g}(|s - t_k| + \|\eta(s) - \eta_k\|) \\ &\le L_{\nabla g}(1 + \|v_k\|)(s - t_k) \le (C(M+1)+1) L_{\nabla g} h \end{aligned}$$

and continue the inequality with

$$g(t, \eta(t)) \le -\alpha(t - t_k) + (C(M+1)+1)^2 L_{\nabla g} h (t - t_k) \le -\frac{\alpha}{2}(t - t_k) \le 0.$$

Therefore, $\eta(t) \in \Theta(t)$ is close to η_k with

$$\operatorname{dist}(\eta_k, \Theta(t)) \le \|\eta_k - \eta(t)\| = (t - t_k)\|v_k\| \le C(M+1)(t - t_k).$$

(ii) If $\binom{t_k}{\eta_k} \notin B_\mu(\operatorname{graph} \partial\Theta(\cdot))$, then $\binom{t_k}{\eta_k} \notin \operatorname{graph} \partial B_\mu(\Theta(\cdot))$ and $\operatorname{dist}(\eta_k, \partial\Theta(t_k))$ is greater than μ. Let us assume that $g(t, \eta_k) > 0$. With the continuous function $\varphi(s) := g(s, \eta_k)$ on $[t_k, t_{k+1}]$, we will soon arrive at a contradiction. Since the inequalities

$$\begin{aligned} \varphi(t_k) &= g(t_k, \eta_k) < 0, \\ \varphi(t) &= g(t, \eta_k) > 0 \end{aligned}$$

hold, there exists $\bar{t} \in (t_k, t) \subset (t_k, t_{k+1}]$ with $\varphi(\bar{t}) = 0$. Then $g(\bar{t}, \eta_k) = 0$ and $\eta_k \in \partial\Theta(\bar{t})$ such that $\binom{\bar{t}}{\eta_k} \in \operatorname{graph} \partial\Theta(\cdot)$. The following inequality shows the contradiction:

$$\operatorname{dist}\left(\binom{t_k}{\eta_k}, \operatorname{graph} \partial\Theta(\cdot)\right) \le \left\|\binom{t_k}{\eta_k} - \binom{\bar{t}}{\eta_k}\right\| = |\bar{t} - t_k| \le h \le \mu.$$

Hence, the assumption was wrong; consequently, it holds $g(t, \eta_k) \le 0$ so that $\eta_k \in \Theta(t)$.

In both cases (i) and (ii), $\operatorname{dist}(\eta_k, \Theta(t)) \le C(M+1)(t - t_k)$. Using (4.1), we get

$$\operatorname{dist}(\widetilde{y}(t), \Theta(t)) \le L_y |t - t_k| + \widetilde{C}_1 h + C(M+1)(t - t_k) \le (C(M+1) + \widetilde{C}_1 + L_y) h,$$

where L_y is the uniform Lipschitz constant from Lemma 2.5. Therefore, a solution $y(\cdot) \in \mathcal{Y}^\Theta[T, t_0, y_0]$ exists by Theorem 3.1 with

$$\sup_{t \in I} \|y(t) - \widetilde{y}(t)\| \le \widetilde{C}_3 h.$$

Hence,

$$\begin{aligned} \|\eta_k - y(t_k)\| &\le \|\eta_k - \widetilde{y}(t_k)\| + \|\widetilde{y}(t_k) - y(t_k)\| \le (\widetilde{C}_1 + \widetilde{C}_3) h, \\ \operatorname{dist}_\infty(\eta, \mathcal{Y}^\Theta[T, t_0, y_0]) &\le (\widetilde{C}_1 + \widetilde{C}_3) h. \qquad \square \end{aligned}$$

5. Example. The dynamical system, underlying the following two test examples, is due to Petar Kenderov. It serves as a model problem for the illustration of first order convergence. We restrict ourselves to the visualization of the convergence of reachable sets. The visualization of the convergence of the whole discrete solution sets would require much more space and the choice of more appropriate data structures.

Naturally, the realization of the set-valued Euler's method (1.4)–(1.5) on a computer amounts to an additional perturbation of the set-valued right-hand side of order 1 and an evaluation of the set union with a local error of order 2 (with respect to Hausdorff distance, uniformly in $t \in I$); for computational details, cf. [6].

Example 5.1. Consider the differential inclusion

$$
\begin{aligned}
&y'(t) \in F(t, y(t)) = \{Ay(t) + uBy(t) \in \mathbb{R}^2 \,:\, 0 \le u \le 1\} \quad \text{(a.e. } t \in [0, 8]),\\
&y(t) \in \Theta := \{y \in \mathbb{R}^2 \,:\, g(y) \le 0\},\\
&y(0) = y_0 = \tbinom{2}{2},
\end{aligned}
$$

where

$$
A = \begin{pmatrix} \sigma^2 - 1 & \sigma\sqrt{1-\sigma^2} \\ -\sigma\sqrt{1-\sigma^2} & \sigma^2 - 1 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & -2\sigma\sqrt{1-\sigma^2} \\ 2\sigma\sqrt{1-\sigma^2} & 0 \end{pmatrix},
$$

$$
g(y) := -\frac{1}{2}(y_1 - 2)^2 + 2 - y_2, \qquad y = \tbinom{y_1}{y_2},
$$

and $\sigma \in (0, 1)$ is a fixed parameter.

The reachable set for the unconstrained case can be expressed by representing its points with polar coordinates,

$$
\begin{aligned}
\mathcal{R}\Big(t, t_0, r_0 \tbinom{\cos(\phi_0)}{\sin(\phi_0)}\Big) = \Big\{ r(t) \tbinom{\cos(\phi(t))}{\sin(\phi(t))} \,:\, & r(t) = r_0 e^{(\sigma^2-1)t},\\
& \phi(t) = \phi_0 + \sigma\sqrt{1-\sigma^2}(2u-1)t, \quad 0 \le u \le 1 \Big\},
\end{aligned}
$$

where the initial point y_0 has polar coordinates $(r_0, \phi_0) = (2\sqrt{2}, \frac{\pi}{4})$. Further on, we fix $\sigma = \frac{9}{10}$.

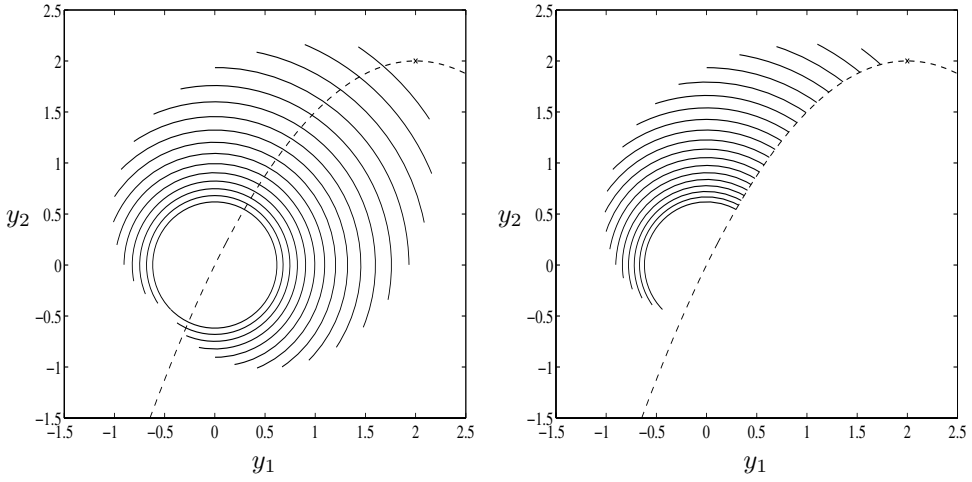


FIG. 5.1. *Reachable sets for different end times t (without, resp., with state constraints).*

In Figure 5.1 (left picture), the exact reachable sets for the unconstrained problem with varying end time $t_i = i \cdot \frac{1}{2}$, $i = 0, \ldots, 16$, and the boundary of the (quadratic) state constraint (dotted line), are illustrated. For $t = 0$, the starting set is just the upper right point in this figure (marked by the cross); for increasing time t the reachable set moves to the lower left of the figure and the two ends of the arcs approach each other. Approximately for $t \ge 8$, the two end points of the arc will overlap and the reachable sets form the boundary of a circle. In the right picture of Figure 5.1, the reachable sets for the state-constrained problem are visualized for the same times. In contrast to [6, Example 5.2.2] with a linear constraint, the reachable set cannot be gained by the intersection $\mathcal{R}(t, t_0, Y_0) \cap \Theta$, as the comparison of both pictures shows.

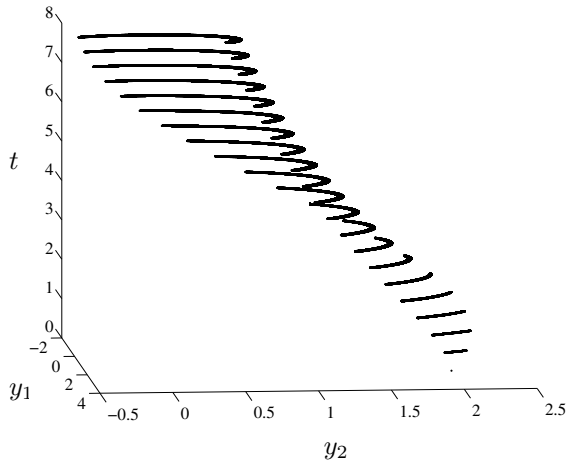


FIG. 5.2. *Integral funnel of Euler's method with* $N = 240$ *and state constraints.*

For $t \geq 7.2$, the small part of the circle that moves out of the interior of Θ (everything below the quadratic function) originates from points that were already cut off by the quadratic state constraint at an earlier time.

Figure 5.2 shows the integral funnel with state constraints. This figure was calculated with set-valued Euler's method for $N = 240$ on the interval $[0, 8]$ (cf. [6, Chapter 4] for details on the implementation of Euler's method for the approximation of nonlinear differential inclusions).

Let us check whether Theorem 4.2 for state-constrained Euler's method can be applied. Observe that $F(t, y) = \{f(t, y, u) : u \in [0, 1]\}$ with $f(t, y, u) = Ay + uBy$ is Lipschitz with respect to (t, y) and has nonempty, compact, convex images. Clearly, (H1) and (C1) are also fulfilled. Furthermore,

$$\langle \nabla g(y)^\top, v \rangle = -(\sigma^2 - 1) \cdot \frac{1}{2} y_1^2 + \sigma \sqrt{1 - \sigma^2} \cdot (1 - 2u) \cdot \frac{1}{2} y_1 \cdot (y_1^2 - 6y_1 + 10)$$

for all $y \in \partial\Theta$ and $v = f(t, y, u)$.

For $y_1 < 0$, the choice of $u = 0$ yields

$$\langle \nabla g(y)^\top, v \rangle = \frac{1}{2} y_1 \cdot \sigma \sqrt{1 - \sigma^2} \cdot \underbrace{\left(y_1^2 - \left(6 - \frac{1}{\sigma} \sqrt{1 - \sigma^2} \right) y_1 + 10 \right)}_{=: h(y_1)}.$$

A discussion of the function h shows $h(y_1) \geq (y_1 - 4)^2 - 6 \geq 10$ so that the scalar product is less than zero.

For $y_1 \in (0, \frac{5}{2}]$, $u = 1$ is chosen such that

$$\langle \nabla g(y)^\top, v \rangle = -\frac{1}{2} y_1 \cdot \sigma \sqrt{1 - \sigma^2} \cdot \left(y_1^2 - \left(6 + \frac{1}{\sigma} \sqrt{1 - \sigma^2} \right) y_1 + 10 \right).$$

The quadratic function in this term could be strictly estimated from below by the function $\widetilde{h}(y_1) = y_1^2 - \frac{13}{2} y_1 + 10$ which is strictly decreasing and is not less than $\widetilde{h}(\frac{5}{2}) = 0$. Hence, the scalar product is also negative.

Let us note that the final reachable set is a circle avoiding the origin; cf. Figure 5.1. Therefore, all discrete reachable sets for small step-sizes have a positive distance to

TABLE 5.1
Estimated order of convergence for $T = 0.5$ (state-constrained problem).

N	Estimated Hausdorff distance from the reference set	Difference to Ch^p
16	0.0897488	-1.1E-02
32	0.0280925	9.2E-03
64	0.0182812	-6.6E-04
128	0.0104471	-2.1E-03
256	0.0036226	3.2E-04
512	0.0018178	4.8E-05

TABLE 5.2
Estimated order of convergence for $T = 7.5$ (state-constrained problem).

N	Estimated Hausdorff distance from the reference set	Difference to Ch^p
16	0.3275207	1.1E-02
32	0.1842108	-7.3E-03
64	0.0923952	-1.2E-04
128	0.0483326	-2.0E-04
256	0.0250892	2.0E-05
512	0.0129622	1.4E-04

the origin so that on a compactum containing all Euler solutions and near to the boundary of Θ we have a positive distance to the origin. A compactness argument yields therefore the validity of (C2). Hence, order of convergence 1 with respect to the step-size h holds by Theorem 4.2.

For the state-constrained case, Tables 5.1 and 5.2 visualize the order of convergence for the approximation of the reachable set $\mathcal{R}(0.5, 0, \binom{2}{2})$, resp., $\mathcal{R}(7.5, 0, \binom{2}{2})$. The tables are calculated by using the theoretical reachable set as reference set. Based on these data, a least squares problem with the function $\log(Ch^p)$ with unknowns $C, p \geq 0$ yields the values $p = 1.0800$ and $C = 1.1812$, resp., $p = 0.9388$ and $C = 1.9156$. The estimated order of convergence for $T = 7.5$ is slightly worse than for $T = 0.5$ due to possible increasing rounding errors.

In Figure 5.3 the difference between the discrete reachable sets generated by Euler's method (gray shaded sets) and the theoretical one (arc with black solid line, almost included in the gray sets) is depicted. The pictures show the approximations of the reachable set with state constraints at time $t = 7.5$ for several numbers N of subintervals: $N = 16$ (upper left), 32 (upper right), 64 (lower left), and 128 (lower right).

Example 5.2. Consider the modified Example 5.1 in which the state constraint is now time-dependent, i.e.,

$$y(t) \in \Theta(t) := \{y \in \mathbb{R}^2 \,:\, g(t,y) \leq 0\},$$
$$g(t,y) := -\frac{1}{4} \cdot \left(2 - \frac{t^2}{64}\right) \cdot (y_1 - 2)^2 + \left(2 - \frac{t^2}{64}\right) - y_2, \quad y = \binom{y_1}{y_2}.$$

Observe that $g(0, y)$ equals the time-independent state constraint in Example 5.1. From Figure 5.4, it is clear that in the case of time-dependent constraints (right picture), the reachable sets are bigger than in the time-independent case (left picture). This figure shows the discrete reachable sets for the constrained problem at the times $t \in \{0, \frac{1}{2}, 1, \frac{3}{2}, 2, 3, 4, 6, 8\}$. For these times, the boundaries of the state constraints $g(t, \cdot) = 0$ are depicted in the right picture with dotted lines.

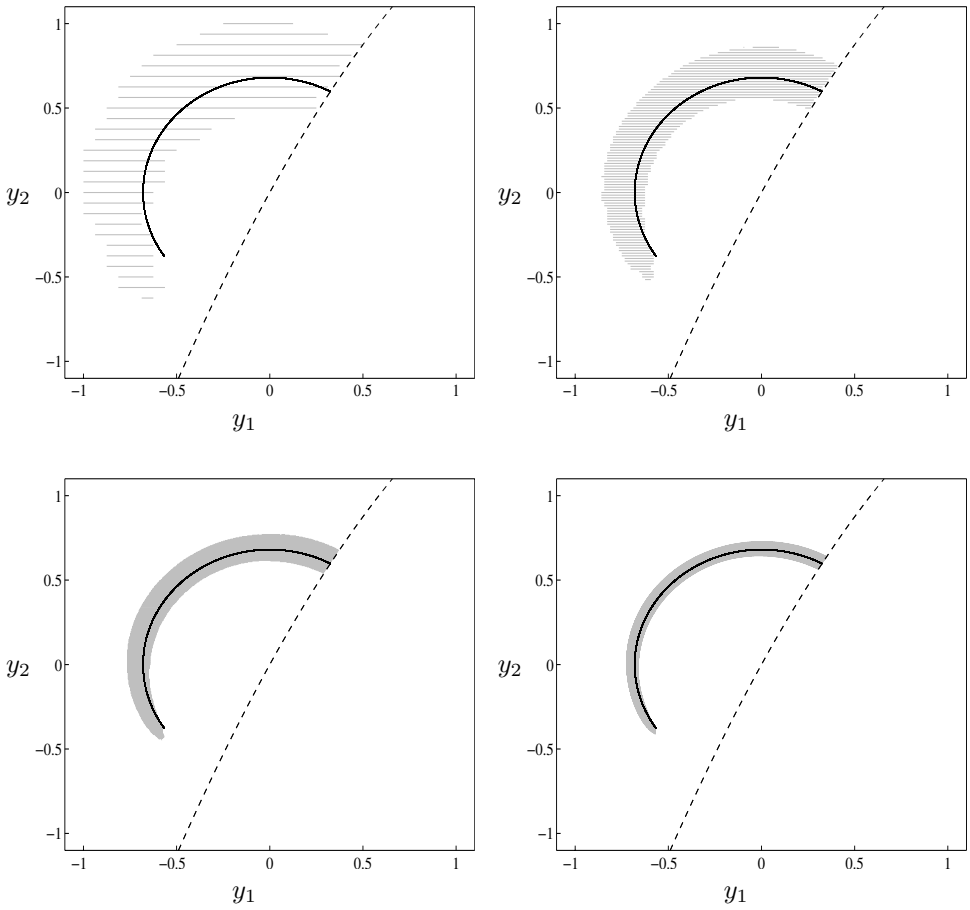


FIG. 5.3. *Discrete reachable sets for* $T = 7.5$ *and various step-sizes* $N = 16, 32, 64, 128$.

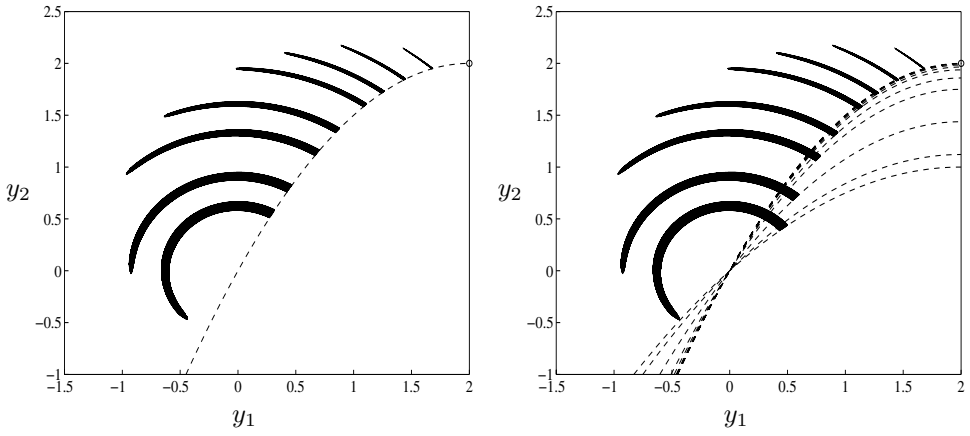


FIG. 5.4. *Discrete reachable sets from Euler's method with* $N = 128$ *for both examples.*

TABLE 5.3
Estimated order of convergence for $T = 7.5$ (time-dependent state-constrained problem).

N	Estimated Hausdorff distance from the reference set	Difference to Ch^p
16	0.3371631	7.2E-03
32	0.1842111	-5.1E-03
64	0.0931844	-4.1E-05
128	0.0483323	1.1E-04
256	0.0250866	1.1E-04
512	0.0130925	1.2E-05

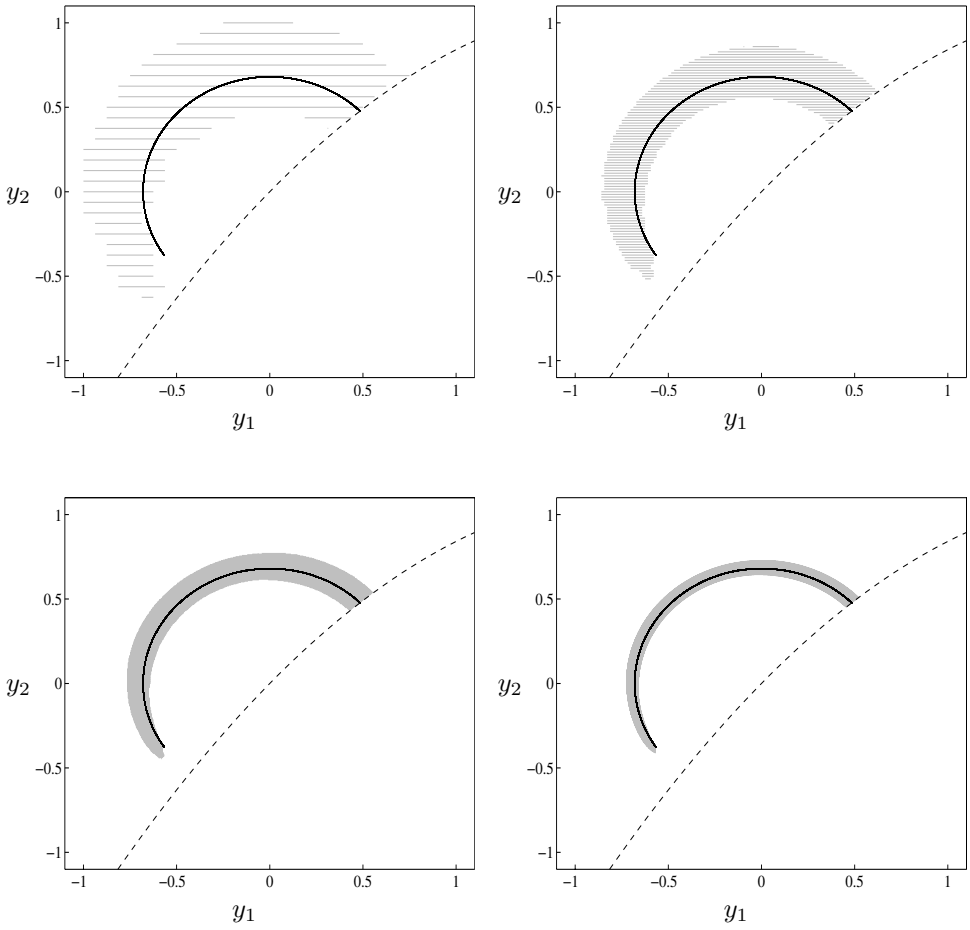


FIG. 5.5. *Discrete reachable sets for $T = 7.5$ and various step-sizes $N = 16, 32, 64, 128$.*

With considerably more effort, it is even possible to show the validity of (C2) by choosing the same values u depending on the sign of y_1 as in Example 5.1.

Table 5.3 is created for the time $T = 7.5$ similarly to the tables for the previous example, but includes the data for the time-dependent state constraint. A least squares approximation with $\log(Ch^p)$ yields the values $p = 0.9431$ and $C = 1.9387$.

Figure 5.5 visualizes how the discrete reachable sets (gray shaded) generated by Euler's method approximate the theoretical reachable set.

REFERENCES

[1] J.-P. Aubin, *Viability Theory*, Systems Control Found. Appl., Birkhäuser Boston, Boston, MA, 1991.

[2] J.-P. Aubin and A. Cellina, *Differential Inclusions*, Grundlehren Math. Wiss. 264, Springer-Verlag, Berlin, 1984.

[3] J.-P. Aubin and H. Frankowska, *Set-Valued Analysis*, Systems Control Found. Appl. 2, Birkhäuser Boston, Boston, MA, 1990.

[4] J. Bastien and M. Schatzman, *Numerical precision for differential inclusions with uniqueness*, M2AN Math. Model. Numer. Anal., 36 (2002), pp. 427–460.

[5] H. Brézis, *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*, North–Holland Math. Stud. 5, Notas Mat. 50, North-Holland, Amsterdam, London; American Elsevier, New York, 1973.

[6] I. A. Chahma, *Set-valued discrete approximation of state-constrained differential inclusions*, Bayreuth. Math. Schr., 67 (2003), pp. 3–162.

[7] F. H. Clarke, L. Rifford, and R. J. Stern, *Feedback in state constrained optimal control*, ESAIM Control Optim. Calc. Var., 7 (2002), pp. 97–133.

[8] F. H. Clarke and R. J. Stern, *State constrained feedback stabilization*, SIAM J. Control Optim., 42 (2003), pp. 422–441.

[9] K. Deimling, *Multivalued Differential Equations*, de Gruyter Ser. Nonlinear Anal. Appl. 1, Walter de Gruyter, Berlin, New York, 1992.

[10] A. Dontchev and F. Lempio, *Difference methods for differential inclusions: A survey*, SIAM Rev., 34 (1992), pp. 263–294.

[11] A. L. Dontchev and E. M. Farkhi, *Error estimates for discretized differential inclusions*, Computing, 41 (1989), pp. 349–358.

[12] A. L. Dontchev and W. W. Hager, *The Euler approximation in state constrained optimal control*, Math. Comp., 70 (2001), pp. 173–203.

[13] A. L. Dontchev, W. W. Hager, and K. Malanowski, *Error bounds for Euler approximation of a state and control constrained optimal control problem*, Numer. Funct. Anal. Optim., 21 (2000), pp. 653–682.

[14] A. F. Filippov, *Differential Equations with Discontinuous Righthand Sides*, Math. Appl. (Soviet Ser.), Kluwer Academic Publishers, Dordrecht, 1988.

[15] F. Forcellini and F. Rampazzo, *On nonconvex differential inclusions whose state is constrained in the closure of an open set. Applications to dynamic programming*, Differential Integral Equations, 12 (1999), pp. 471–497.

[16] H. Frankowska, S. Plaskacz, and T. Rzeżuchowski, *Measurable viability theorems and the Hamilton-Jacobi-Bellman equation*, J. Differential Equations, 116 (1995), pp. 265–305.

[17] H. Frankowska and F. Rampazzo, *Filippov's and Filippov-Wazewski's theorems on closed domains*, J. Differential Equations, 161 (2000), pp. 449–478.

[18] H. Frankowska and R. B. Vinter, *Existence of neighboring feasible trajectories: Applications to dynamic programming for state-constrained optimal control problems*, J. Optim. Theory Appl., 104 (2000), pp. 21–40.

[19] F. Lempio, *Difference methods for differential inclusions*, in Modern Methods of Optimization. Proceedings of a Summer School at the Schloß Thurnau of the University of Bayreuth (Germany), FRG, 1990, Lecture Notes in Econom. and Math. Systems 378, Springer-Verlag, Berlin, 1992, pp. 236–273.

[20] F. Lempio, *Euler's method revisited*, Proc. Steklov Inst. Math., 211 (1995), pp. 429–449.

[21] F. Lempio and D. Silin, *Generalized differential equations with strongly one-sided Lipschitzian right-hand side*, Differential Equations, 32 (1996), pp. 1485–1491.

[22] H. M. Soner, *Optimal control with state-space constraint* I, SIAM J. Control Optim., 24 (1986), pp. 552–561.

[23] R. J. Stern, *Characterization of the state constrained minimal time function*, SIAM J. Control Optim., 43 (2004), pp. 697–707.

[24] V. M. Veliov, *Second order discrete approximations to strongly convex differential inclusions*, Systems Control Lett., 13 (1989), pp. 263–269.

[25] V. Veliov, *Second-order discrete approximation to linear differential inclusions*, SIAM J. Numer. Anal., 29 (1992), pp. 439–451.

SIAM J. OPTIM.
Vol. 18, No. 3, pp. 1027–1045

SOME NONLINEAR MAPS AND RENORMINGS OF BANACH SPACES*

S. LAJARA†, A. J. PALLARÉS‡, AND S. TROYANSKI‡

Abstract. We consider two classes of nonlinear maps between normed spaces which are relevant for the study of Banach spaces that admit equivalent locally uniformly rotund and midpoint locally uniformly rotund norms. It turns out to be a useful characterization of such maps in terms of optimization, topology, and probability. Using these characterizations we obtain some renorming results.

Key words. σ-midpoint continuous map, σ-slicely continuous map, midpoint locally uniformly rotund norm

AMS subject classifications. Primary, 46B20; Secondary, 54E99, 54H05

DOI. 10.1137/060656280

1. Introduction.

Renorming a Banach space X consists in finding equivalent norms on X with good geometrical properties of convexity or differentiability, as close as possible to those of the euclidean norms in finite dimensional Banach spaces.

Among the notions related to the convexity we have that of a locally uniformly rotund norm, which plays an important role in the geometry of Banach spaces as well as in optimization theory (see, e.g., [4, 5, 17, 18]). A normed space X (or a norm $\|\cdot\|$ on X) is said to be *locally uniformly rotund* (LUR) if, for every $x \in X$ and every sequence $(x_n)_n \subset X$ such that $\lim_n \|x_n\| = \|x\|$ and $\lim_n \|x_n + x\| = 2\|x\|$, we have $\lim_n \|x_n - x\| = 0$.

The class of normed spaces that admit an equivalent LUR norm was characterized in [12] (see also [16]), involving the notion of ϵ-denting point. Recall that an element x of a set K in a normed space X is an *ϵ-denting point* of K if there exists an open half space $H \subset X$ (i.e., a set of the form $f^{-1}(a, \infty)$ with $f \in X^*$ and $a \in \mathbb{R}$) such that $x \in H$ and $\operatorname{diam}(H \cap K) < \epsilon$. The set $H \cap K$ is called a *slice* of K.

THEOREM 1 (see [12, Main Theorem]). *A normed space X admits an equivalent LUR norm if and only if for every $\epsilon > 0$ we can write*

$$X = \bigcup_{n \in \mathbb{N}} X_{n,\epsilon}$$

in such a way that every $x \in X_{n,\epsilon}$ is an ϵ-denting point of the set $X_{n,\epsilon}$.

It is rather difficult to apply directly the above theorem to get LUR norms in concrete normed spaces. In [14] a new class of maps was introduced and studied, the σ-slicely continuous maps. Using this class of maps, a transfer technique for

*Received by the editors April 4, 2006; accepted for publication (in revised form) May 28, 2007; published electronically October 4, 2007. This research was supported by MCYT MTM 2005-08379 and Fundación Séneca 00690/PI/04 CARM.

http://www.siam.org/journals/siopt/18-3/65628.html

†Departamento de Matemáticas, Escuela Politécnica Superior, Universidad de Castilla La Mancha, Campus Universitario, 02071 Albacete, Spain (sebastian.lajara@uclm.es). This author was supported by JCCM PAI-05-034.

‡Departamento de Matemáticas, Universidad de Murcia, Campus de Espinardo, 30100 Murcia, Spain (apall@um.es, stroya@um.es). The third author was supported by a project of the Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, and grant MM-1401/04 of the Bulgarian NFR.

LUR renorming was established. It turns out that this class of maps admits characterizations in terms of optimization (subdifferentials), probability (expectations and distributions of random variables), and topology (σ-discrete families of sets).

As a variant of the LUR property we have the notion of MLUR norm. A normed space X (or a norm $\|\cdot\|$ on X) is said to be *midpoint locally uniformly rotund* (MLUR) if, for every $x \in X$ and every sequence $(x_n)_n \subset X$ such that $\|x_n + x\| \to \|x\|$ and $\|x_n - x\| \to \|x\|$, we have $\|x_n\| \to 0$.

It is well known that the MLUR property lies between local uniform rotundity and strict convexity. In the paper [7], devoted to the renorming of spaces of continuous functions on trees, Haydon provided the first example of MLUR space with no equivalent LUR renorming. On the other hand, the space ℓ_∞ (which admits an equivalent dual strictly convex norm being the dual of a separable space) does not have any equivalent MLUR renorming (see, e.g., [2] and [8], where it was shown that there is not any equivalent MLUR renorming on ℓ_∞ using a variant of the argument in [11] about the impossibility of LUR renormability of that space).

The class of MLUR renormable spaces was characterized in a linear topological way in [13], using the concept of ϵ-strongly extreme point, introduced in [9]. The following definition provides an equivalent formulation of this concept.

DEFINITION 2. *Let K be a subset of a normed space X, and let $\epsilon, \delta > 0$. An element $x \in K$ is said to be an (ϵ, δ)-strongly extreme point of K if*

$$\|u - x\| \le \epsilon \quad \text{and} \quad \|v - x\| \le \epsilon \quad \text{whenever} \quad u, v \in K \quad \text{and} \quad \left\| x - \frac{u+v}{2} \right\| < \delta.$$

The point x is called an ϵ-strongly extreme point *of K if there exists $\delta > 0$ such that x is an (ϵ, δ)-strongly extreme point of K.*

It is easy to see that every ϵ-denting point is a 2ϵ-strongly extreme point. It is also well known that a normed space is MLUR if and only if every element of its unit sphere is an ϵ-strongly extreme point of the unit ball for each $\epsilon > 0$. The aforementioned characterization of MLUR renormability is given by the following theorem.

THEOREM 3 (see [13, Theorem 1]). *A normed space X admits an equivalent MLUR norm if and only if for every $\epsilon > 0$ we can write*

$$X = \bigcup_{n=1}^{\infty} X_{n,\epsilon}$$

in such a way that each $x \in X_{n,\epsilon}$ is an ϵ-strongly extreme point of the set $\mathrm{co}(X_{n,\epsilon})$, *where* $\mathrm{co}(A)$ *denotes the convex hull of A.*

In this paper we provide some more characterizations of the σ-slicely continuous maps. Also, motivated by the notion of ϵ-strongly extreme point and the above characterization of MLUR renormability, we introduce and study a new class of nonlinear maps between normed spaces that we call σ-midpoint continuous maps. This class includes that of the σ-slicely continuous maps and is relevant for the study of normed spaces that admit equivalent MLUR norms. In terms of this new class we obtain some results which cover all the known cases of MLUR renorming.

2. σ-slicely continuous and σ-midpoint continuous maps. As we mentioned in the introduction, in the recent memoir [14], a nonlinear transfer technique for LUR renormability has been developed. The following notion, introduced there, plays the main role in this technique.

DEFINITION 4. *Let X and Y be normed spaces, and let A be a subset of X. A map $\Phi : A \longrightarrow Y$ is said to be σ-slicely continuous if, for every $\epsilon > 0$, we may write*

$$A = \bigcup_{n\in\mathbb{N}} A_{n,\epsilon}$$

in such a way that for every $x \in A_{n,\epsilon}$ *there exists an open half space* $H_x \subset X$ *such that* $x \in H_x$ *and* $\operatorname{diam}\Phi(A_{n,\epsilon} \cap H_x) < \epsilon$.

This concept can be regarded as the countable covering counterpart of the notion of slice continuity. Recall that if A is a subset of X, then a map $\Phi : A \longrightarrow Y$ is said to be *slicely continuous* at a point $x \in A$ if there exists an open half space $H_x \subset X$ containing x such that $\operatorname{diam}\Phi(A \cap H_x) < \epsilon$.

The characterization of LUR renormable spaces given in Theorem 1 shows that a normed space X is LUR renormable if and only if the identity operator $Id : X \longrightarrow X$ is σ-slicely continuous.

Let us note that if Y is a separable normed space, then every map $\Phi : X \longrightarrow Y$ is σ-slicely continuous. Indeed, for any $\epsilon > 0$ we can write $Y = \bigcup_{n\in\mathbb{N}} B_n$, where each B_n is a ball of diameter less than ϵ. The sequence $\{\Phi^{-1}(B_n)\}_n$ is then a countable covering of X that satisfies the required properties. Therefore, the natural framework for the study of the class of σ-slicely continuous maps is that of nonseparable spaces.

In [14], many examples of maps in this class are given. Let us mention that it includes, among others, bounded linear operators whose domain or range is an LUR renormable space and positive convex maps with values in an LUR lattice. Recall that a map Φ from a normed space X into a Banach lattice Y is said to be positive and convex if for every $x_1, x_2 \in X$ and $\lambda, \mu \geq 0$ such that $\lambda + \mu = 1$ we have

$$0 \leq \Phi(\lambda x_1 + \mu x_2) \leq \lambda\Phi x_1 + \mu\Phi x_2.$$

Also, in [14], the σ-slicely continuous maps are characterized using the notion of ϵ-subdifferential, used by Asplund and Rockafellar [3] in nonlinear analysis. Recall that if A is a subset of a linear topological space X and ϕ is a map from A into $\mathbb{R}$, the ϵ-subdifferential of ϕ at a point $x \in C \subseteq A$ is the set

$$\partial_\epsilon\phi(x|C) = \{f \in X^* : f(y) \geq \phi(x) + f(y - x) - \epsilon \quad \forall y \in C\}$$

and the subdifferential of ϕ at x is $\partial\phi(x|C) = \bigcap_{\epsilon>0} \partial_\epsilon\phi(x|C)$.

In the next theorem we complete the characterization of σ-slicely continuous maps in a probabilistic way. Before establishing that theorem, we make some comments about the notations and background from measure theory to be used in its statement and proof (as well as in other definitions and results in this work).

We shall consider a probability space (Ω, Σ, P), where Σ is a σ-algebra of subsets of Ω. A function $U : \Omega \longrightarrow X$ is said to be *simple* if there exist $x_1, \ldots, x_m \in X$ and $E_1, \ldots, E_m \in \Sigma$ such that $U = \sum_{k=1}^m x_k \mathbb{1}_{E_k}$, where $\mathbb{1}_{E_k}$ denotes the characteristic function of the set E_k. A function $U : \Omega \longrightarrow X$ is called *strongly measurable* if there exists a sequence of simple functions $U_n : \Omega \longrightarrow X$ such that $\lim_n \|U - U_n\| = 0$ almost everywhere. A strongly measurable function $U : \Omega \longrightarrow X$ is said to be *Bochner integrable* whenever there is a sequence of simple functions U_n such that $\lim_n \int \|U - U_n\| dP = 0$. In this case, the integral with respect to a set $E \in \Sigma$ is $\lim_n \int_E U_n dP$, where the integral for simple functions is defined in the usual way. The symbol $L^1(P, X)$ stands for the Banach space of all Bochner integrable random variables $U : \Omega \longrightarrow X$ endowed with the norm $\|\cdot\|_1$. We refer the reader to the survey [6] for a complete study of the Bochner integral.

We shall consider random variables $U : \Omega \longrightarrow X$, which we always assume to be Bochner integrable functions. We write $\mathbb{E}(U) = \int_\Omega U\, dP$ for the expectation of U

with respect to the probability P, and $\mathbb{E}_B V = (\int_B V dP)/P(B)$ for its conditional expectation with respect to a measurable set $B \subset \Omega$ such that $P(B) > 0$.

Since every Bochner integrable function $U : \Omega \longrightarrow X$ can be approximated in the $L^1(P, X)$-norm by a sequence of functions taking finitely many values in $\mathrm{co}(U(\Omega))$ (see, e.g., [6, Chapters II.2.8 and V.2.2]), for our purposes it is enough to consider random variables U which take finitely many values $\{u_i\}_{i\in I}$ with probabilities $\{p_i\}_{i\in I}$. In this case, clearly $\mathbb{E}(U) = \sum_{i\in I} p_i u_i$. However, in order to simplify our presentation we shall consider general Bochner integrable random variables.

In the definition of σ-slicely continuous map, it is not considered any additional property ensuring the strong measurability of the map $\Phi \circ U$ whenever U is a strongly measurable function. However, in most of the examples and applications, Φ is norm continuous, and when we are able to ensure that $\Phi \circ U$ is bounded, we also have that this function is Bochner integrable.

THEOREM 5. *Let X and Y be normed spaces, A be a subset of X, and Φ be a map from A into Y. The following conditions are equivalent:*

(a) *Φ is σ-slicely continuous;*
(b) *for every $\epsilon > 0$ we can write $A = \bigcup_{n\in\mathbb{N}} A_{n,\epsilon}$ in such a way that for every $x \in A_{n,\epsilon}$ and every 1-Lipschitzian function $g : \Phi(A_{n,\epsilon}) \longrightarrow \mathbb{R}$ we have $\partial_\epsilon g \circ \Phi(x|A_{n,\epsilon}) \neq \emptyset$;*
(c) *for every $\epsilon > 0$ and every $\xi > 0$, we can write $A = \bigcup_{n\in\mathbb{N}} A_{n,\epsilon,\xi}$ in such a way that for every $x \in A_{n,\epsilon,\xi}$ there exists a number $\delta > 0$ with the property that if $U : (\Omega, \Sigma, P) \to A_{n,\epsilon,\xi}$ is a random variable such that ΦU is also a random variable and $\|x - \mathbb{E}(U)\| < \delta$, then $P(\{\|\Phi x - \Phi U\| > \epsilon\}) < \xi$;*
(d) *for every $\epsilon > 0$ we can write $A = \bigcup_{n\in\mathbb{N}} A_{n,\epsilon}$ in such a way that for every $x \in A_{n,\epsilon}$ there exists a number $\delta > 0$ with the property that if $U : (\Omega, \Sigma, P) \to A_{n,\epsilon}$ is a random variable such that ΦU is also a random variable and $\|x - \mathbb{E}(U)\| < \delta$, then $\mathbb{E}(\|\Phi x - \Phi U\|) < \epsilon$.*

In order to complete the proof of this theorem we need the following lemma, which will be used in other results through this paper. In its proof, we take advantage of some technical ideas from [13, Proposition 1] and [14, Lemma 4.21]. Let us observe that if U is a random variable and $f \in X^*$, $f \circ U$ is a measurable function, and that if $H = f^{-1}(\alpha, \infty)$ is an open half space, then the set $\{U \in H\} = \{\omega \in \Omega : f(U\omega) > \alpha\}$ is measurable.

LEMMA 6. *Let A be a subset of a normed space X. Assume that for each $x \in A$ we have chosen an open half space $H_x \subset X$ such that $x \in H_x$. Then for each fixed $\xi > 0$ we can write $A = \bigcup_{n\in\mathbb{N}} A_n$ in such a way that for every $n \in \mathbb{N}$ and every $x \in A_n$ there exist an open half space H'_x, a number $\eta_x > 0$, and an element $w_x \in A$ with the following properties:*

(a) *$x \in H'_x \cap A_n \subset H_{w_x} \cap A$;*
(b) *if U is a random variable on a probability space (Ω, Σ, P) with values in A_n such that $\|x - \mathbb{E}(U)\| < \eta_x$, then $P(\{U \notin H'_x\}) < \xi$.*

Proof. We can assume without loss of generality that the set A is bounded. So,

$$\sup\{f(x) : x \in A,\ f \in X^*,\ \|f\| = 1\} < \infty. \tag{1}$$

Let us fix, for each $x \in X$, a norm-one functional $f_x \in X^*$ and rational numbers q_x and r_x (with $r_x > 0$) such that $\{z \in X : f_x(z) > q_x\} \subseteq H_x$ and

$$f_x(x) > q_x + r_x.$$

We make an initial decomposition of the set A by defining, for each $q \in \mathbb{Q}$,

$$A_q = \{x \in A : q_x = q\}.$$

Now for each pair $r, s \in \mathbb{Q}^+$ we write

$$A_{q,r,s} = \left\{ x \in A_q : s < \xi r_x/2,\ q + r < \sup_{w \in A_q} f_w(x) < q + r + s \right\}.$$

Thanks to (1) we have

$$A = \bigcup_{q \in \mathbb{Q},\, r,s \in \mathbb{Q}^+} A_{q,r,s}.$$

Let us fix $x \in A_{q,r,s}$, let us take an element $w_x \in A_q$ such that

$$f_{w_x}(x) > q + r,$$

and let

$$H'_x = \{z \in X : f_{w_x}(z) > f_{w_x}(x) - r\}.$$

If $z \in H'_x \cap A_{q,r,s}$, then $f_{w_x}(z) > q = q_{w_x} = q_x$ and

$$f_{w_x}(z) < q + r + s < f_{w_x}(x) + s.$$

The first inequality means that $H'_x \cap A_{q,r,s} \subseteq H_{w_x}$. From the second we deduce that

$$f_{w_x}(x - z) > -s. \tag{2}$$

On the other hand, for each $z \in A_{q,r,s} \setminus H'_x$ we have

$$f_{w_x}(x - z) \geq r. \tag{3}$$

Note also that because of the definition of the set $A_{q,r,s}$ we have

$$r + s > f_x(x) - q > r_x. \tag{4}$$

Now choose a number $0 < \eta_x < \min\{s, 1\}$, and let $U : (\Omega, \Sigma, P) \to A_{q,r,s}$ be a random variable such that $\|x - \mathbb{E}(U)\| < \eta_x$. Then, using inequalities (2), (3), and (4) we obtain

$$\begin{aligned}
\eta_x &> \|\mathbb{E}(x - U)\| \geq \mathbb{E}(f_{w_x}(x - U)) \\
&= P(\{U \notin H'_x\})\, \mathbb{E}_{\{U \notin H'_x\}}(f_{w_x}(x - U)) \\
&\quad + P(\{U \in H'_x\})\, \mathbb{E}_{\{U \in H'_x\}}(f_{w_x}(x - U)) \\
&\geq r P(\{U \notin H'_x\}) - s P(\{U \in H'_x\}) \\
&= (r + s) P(\{U \notin H'_x\}) - s \\
&> r_x P(\{U \notin H'_x\}) - s
\end{aligned}$$

and, consequently,

$$P(\{U \notin H'_x\}) < \frac{s + \eta_x}{r_x} < \frac{2s}{r_x} < \xi. \qquad \square$$

Proof of Theorem 5. We start by proving the implication (a) ⇒ (c). Let us fix $\epsilon, \xi > 0$. Since Φ is σ-slicely continuous we have a countable decomposition of A,

$$A = \bigcup_{n\in\mathbb{N}} A_n,$$

in such a way that for every $x \in A_n$ there is an open half space $H_x \subset X$ such that $x \in H_x$ and

$$\operatorname{diam} \Phi(A_n \cap H_x) < \frac{\epsilon}{2}. \tag{5}$$

By Lemma 6, for each $n \in \mathbb{N}$ we can write

$$A_n = \bigcup_{m\in\mathbb{N}} A_{n,m}$$

in such a way that for every $x \in A_{n,m}$ there exist a number $\eta_x > 0$, an open half space $H'_x \subset X$, and an element $w_x \in A_n$ such that $x \in H'_x \cap A_{n,m} \subseteq H_{w_x} \cap A_n$ and if $U : (\Omega, \Sigma, P) \to A_{n,m}$ is a random variable such that $\|x - \mathbb{E}(U)\| < \eta_x$, then $P(\{U \notin H'_x\}) < \xi$. By (5) it follows that $\{\|\Phi x - \Phi U\| > \epsilon\} \subset \{U \notin H'_x\}$. So (assuming that ΦU is also a random variable) we get $P(\|\Phi x - \Phi U\| > \epsilon) < \xi$, and assertion (c) is proved.

To show that (c) ⇒ (d) we assume that there exists $M > 0$ such that $\|\Phi y\| \le M$ for all $y \in A$. Let us fix $\epsilon > 0$. By hypothesis we can write

$$A = \bigcup_{n\in\mathbb{N}} A_n$$

in such a way that for every $x \in A_n$ there is a number $\delta_x > 0$ with the property that if $U : (\Omega, \Sigma, P) \to A_n$ is a random variable such that ΦU is also a random variable and $\|x - \mathbb{E}(U)\| < \delta_x$, then $P(C) < \frac{\epsilon}{4M}$, where $C = \{\|\Phi x - \Phi U\| > \frac{\epsilon}{2}\}$. Thus,

$$\begin{aligned}\mathbb{E}(\|\Phi x - \Phi U\|) &= P(C)\,\mathbb{E}_C(\|\Phi x - \Phi U\|) + P(\Omega \setminus C)\,\mathbb{E}_{\Omega\setminus C}(\|\Phi x - \Phi U\|)\\ &\le 2MP(C) + \frac{\epsilon}{2}P(\Omega \setminus C) < \epsilon,\end{aligned}$$

as we wanted to show.

Now we prove that (d) ⇒ (a). As before we fix $\epsilon > 0$. Let $\{A_{n,\epsilon}\}_n$ be a covering of A such that for every $x \in A_{n,\epsilon}$ there is a number $\delta_x > 0$ with the property that if $U : (\Omega, \Sigma, P) \to A_{n,\epsilon}$ is a random variable such that ΦU is also a random variable and $\|x - \mathbb{E}(U)\| < \delta_x$, then $\mathbb{E}(\|\Phi x - \Phi U\|) < \epsilon$. We are going to prove that this decomposition also satisfies condition (a).

Take $n \in \mathbb{N}$ and $x \in A_{n,\epsilon}$. Let

$$C = \overline{\operatorname{co}\left(\{z \in A_{n,\epsilon} : \|\Phi z - \Phi x\| \ge \epsilon\}\right)}^{\|\cdot\|}.$$

We claim that $\operatorname{dist}(x, C) \ge \frac{\delta_x}{2}$. Indeed, assuming the contrary, we can find positive numbers $p_1, \dots, p_r$ and vectors $u_1, \dots, u_r \in A_{n,\epsilon}$ such that $\sum_{i=1}^r p_i = 1$,

$$\left\| x - \sum_{i=1}^r p_i u_i \right\| < \delta_x, \tag{6}$$

and

$$\|\Phi x - \Phi u_i\| > \epsilon \quad \forall i = 1, \dots, r. \tag{7}$$

Let U be a random variable which takes the values u_i with probabilities p_i. Using (6) and the hypothesis we get $\mathbb{E}(\|\Phi x - \Phi U\|) = \sum_{i=1}^{r} p_i \|\Phi x - \Phi u_i\| < \epsilon$, which is a contradiction with (7).

So, $\operatorname{dist}(x, C) \geq \frac{\delta_x}{2}$. Since the set C is convex and closed, it follows by Hahn–Banach separation theorem the existence of an open half space $H_x \subset X$ such that $x \in H_x$ and $C \subset X \setminus H_x$. In particular, $\operatorname{diam} \Phi(A_{n,\epsilon} \cap H_x) < 2\epsilon$ and the map Φ is σ-slicely continuous.

The equivalence between (a) and (b) is proved in [13]. □

Let us mention that in [13] are given more characterizations of σ-slicely continuous maps in topological terms.

Now we consider a simple particular case of this class of maps that will be used for establishing one of the main renorming results in the next section.

Definition 7. *Let X be a normed space and Λ be a set. A map $\tau : X \longrightarrow \Lambda$ is said to be σ-slicely constant if we can write*

$$X = \bigcup_{n \in \mathbb{N}} X_n$$

in such a way that for every $x \in X_n$ there is an open half space $H_x \subset X$ such that $x \in H$ and $\tau(z) = \tau(x)$ for every $z \in X_n \cap H_x$.

The concept of a σ-slicely constant map has been introduced in [15]. There, it is proved that every σ-slicely continuous map between two normed spaces can be expressed as a norm pointwise limit of a sequence of σ-slicely constant maps.

The rigidity property of the space $c_0(\Gamma)$ provides examples of σ-slicely constant maps in those normed spaces X that admit a σ-slicely continuous map $\Phi : X \longrightarrow c_0(\Gamma)$.

Proposition 8. *Let A be a subset of a normed space X, let $\Phi : A \longrightarrow c_0(\Gamma)$ be a σ-slicely continuous map, and let $\epsilon > 0$. Then the set valued map $M_{\Phi,\epsilon} : A \longrightarrow 2^{\Gamma}$ defined by the formula*

$$M_{\Phi,\epsilon}(x) := \{\gamma \in \Gamma : |\Phi x(\gamma)| \geq \epsilon\}$$

is σ-slicely constant.

Proof. Set, for each $k \in \mathbb{N}$,

$$A_k = \left\{ x \in A : \sup\{|\Phi x(\gamma)| : \gamma \notin M_{\Phi,\epsilon}(x)\} < \epsilon - \frac{1}{k} \right\}.$$

By the σ-slicely continuity of Φ we can write

$$A_k = \bigcup_{n \in \mathbb{N}} A_{n,k}$$

in such a way that for every $x \in A_{n,k}$ there is an open half space $H_x \subset X$ such that $x \in H_x$ and

$$\operatorname{diam} \Phi\left(H_x \cap A_{n,k}\right) < \frac{1}{k}. \tag{8}$$

It is clear that

$$A = \bigcup_{k,n\in\mathbb{N}} A_{n,k}.$$

Now fix $k, n \in \mathbb{N}$, and let $z \in H_x \cap A_{n,k}$. We claim that $M_{\Phi,\epsilon}(z) \subseteq M_{\Phi,\epsilon}(x)$. Assuming the contrary we can find an element $\beta \in \Gamma$ such that $|\Phi z(\beta)| \geq \epsilon$ and $|\Phi x(\beta)| < \epsilon - \frac{1}{k}$. From these inequalities we get

$$|\Phi z(\beta) - \Phi x(\beta)| \geq |\Phi z(\beta)| - |\Phi x(\beta)| > \epsilon - \left(\epsilon - \frac{1}{k}\right) = \frac{1}{k},$$

which is a contradiction with (8). In a similar way we can deduce that $M_{\Phi,\epsilon}(x) \subseteq M_{\Phi,\epsilon}(z)$, and the proposition is proved. □

Now our aim is to introduce new classes of maps between normed spaces to be useful for the study of MLUR renormings. As may be expected, these classes will be inspired by the notion of strongly extreme point and the covering type characterization of MLUR renormability given by Theorem 3. Before formulating the corresponding concepts it is convenient to prove the following simple fact concerning strongly extreme points.

LEMMA 9. *Let K be a set of a normed space X, let $\epsilon, \delta > 0$, and let $x \in K$. Then the following conditions are equivalent:*

(a) *x is an (ϵ, δ)-strongly extreme point of the set $\mathrm{co}(K)$;*

(b) *if $U : (\Omega, \Sigma, P) \to K$ is a random variable such that $\|x - \mathbb{E}(U)\| < \delta$, then for any measurable subset $C \subset \Omega$ with $P(C) \geq \frac{1}{2}$ we have*

$$\|x - \mathbb{E}_C(U)\| \leq \epsilon.$$

Proof. Assume that x is an (ϵ, δ)-strongly extreme point of $\mathrm{co}(K)$. It is also an (ϵ, δ)-strongly extreme point of $\overline{\mathrm{co}(K)}$. Let $U : (\Omega, \Sigma, P) \to K$ be a random variable such that $\|x - \mathbb{E}(U)\| < \delta$, and let $C \in \Sigma$ with $1 > P(C) = \lambda \geq \frac{1}{2}$. Define $w_1 = \mathbb{E}_C(U)$ and $w_2 = \mathbb{E}_{\Omega\setminus C}(U)$. It is clear that $w_1, w_2 \in \overline{\mathrm{co}(K)}$ and that $\mathbb{E}(U) = \lambda w_1 + (1-\lambda)w_2$. The vector $w_3 = (2\lambda - 1)w_1 + (2 - 2\lambda)w_2$ belongs to the line segment joining w_1 and w_2, in particular $w_3 \in \overline{\mathrm{co}(K)}$. Moreover, $\frac{w_1+w_3}{2} = \mathbb{E}(U)$, and then $\left\|x - \frac{w_1+w_3}{2}\right\| < \delta$. Since x is an (ϵ, δ)-strongly extreme point of $\overline{\mathrm{co}(K)}$ we deduce that $\|w_1 - x\| \leq \epsilon$, and assertion (b) is proved.

For the implication (b) $\Rightarrow$ (a) just consider the midpoint of two vectors u, v in $\mathrm{co}(K)$ as the expectation of a random variable U that takes values in K in such a way that there is a set C with $P(C) = \frac{1}{2}$, $u = \mathbb{E}_C(U)$, and $v = \mathbb{E}_{\Omega\setminus C}(U)$. □

DEFINITION 10. *Let X and Y be normed spaces, and let A be a subset of X. A map $\Phi : A \longrightarrow Y$ is said to be midpoint continuous at a point $x \in A$ if, for every $\epsilon > 0$, there is a number $\delta > 0$ with the following property: if U is a simple random variable on the probability space (Ω, Σ, P) with values in A, and $\|x - \mathbb{E}(U)\| < \delta$, then for every measurable set $C \subseteq \Omega$ such that $P(C) \geq \frac{1}{2}$ we have $\|\Phi x - \mathbb{E}_C(\Phi U)\| \leq \epsilon$.*

Let us mention that the number $\frac{1}{2}$ in the above definition can be replaced with any positive constant less than $\frac{1}{2}$.

According to Lemma 9 it follows that the space X is MLUR if and only if the restriction of the identity operator on X to the unit ball is midpoint continuous on the unit sphere of X. If Φ is bounded and A is convex, we can consider arbitrary random variables in the above definition. This is due to the fact that if U is such a random variable, then U can be approximated pointwise and in the $L^1(P, X)$-norm by

a sequence of simple functions U_n with values in A, and ΦU by ΦU_n simultaneously, as a consequence of the dominated convergence theorem.

It is clear that every midpoint continuous map is continuous. The property of being midpoint continuous is, indeed, much stronger than the simple continuity. In fact, if I is an interval of $\mathbb{R}$ and $\Phi : I \longrightarrow \mathbb{R}$ is a midpoint continuous map, then for each $a, b \in I$ we have $\Phi((a+b)/2) = \Phi(a) = \Phi(b)$, and Φ is constant.

It is not difficult to check directly the midpoint continuity of the identity operator of some concrete spaces at some points. We will present a detailed discussion of this fact in the cases of Hilbert and $C(K)$ spaces, in order to make the reader familiar with the probabilistic techniques that we shall use in other results of this paper.

Let H be a Hilbert space. Fix $x \in H$ with $\|x\| = 1$ and $\epsilon > 0$. Let U be a random variable with values in the unit ball B_H such that $\|x - y\| < \delta$, where $0 < \delta < \min\{1, \epsilon^2/32\}$ and $y = \mathbb{E}(U)$. Bearing in mind that $\mathbb{E}(y - U) = 0$ we get

$$\mathbb{E}\left(\|x - y + U\|^2\right) = \mathbb{E}\left(\|x\|^2 + \|y - U\|^2\right).$$

Using this equality and the fact that P is a probability we have

$$\begin{aligned} &\left(\mathbb{E}(\|x - U\|)\right)^2 \leq \mathbb{E}\left((\|x - y\| + \|y - U\|)^2\right) \\ &\leq 2\,\mathbb{E}\left(\|x - y\|^2 + \|y - U\|^2\right) \leq 2\left(\delta^2 + \mathbb{E}(\|x - y + U\|^2) - \|x\|^2\right) \\ &\leq 2\left(\delta^2 + \mathbb{E}\left((\|x - y\| + \|U\|)^2\right) - 1\right) \leq \left(\delta^2 + (\delta + 1)^2 - 1\right) < 8\delta. \end{aligned} \tag{9}$$

So, if C is a measurable set with $P(C) \geq \frac{1}{2}$, then

$$\|x - \mathbb{E}_C(U)\| \leq \mathbb{E}_C\left(\|x - U\|\right) \leq \frac{1}{P(C)}\,\mathbb{E}\left(\|x - U\|\right) < 2\sqrt{8\delta} < \epsilon. \tag{10}$$

Thus, the restriction of the identity operator to the unit ball of H is midpoint continuous at every element of its unit sphere. Let us note that our argument shows that the identity operator on B_H is, actually, slicely continuous at each point of the unit sphere (see (d) $\Rightarrow$ (a) in Theorem 5).

Now let K be a compact set and x be an extreme point of the unit ball of the space $(C(K), \|\cdot\|_\infty)$. It is well known that $|x(t)| = 1$ for each $t \in K$. Fix $\epsilon > 0$, let U be a random variable with values in $B_{C(K)}$ such that $\|x - \mathbb{E}(U)\|_\infty < \delta$, where $0 < \delta < \min\{1, \epsilon^2/32\}$, and let C be a measurable set such that $P(C) \geq \frac{1}{2}$. For every fixed $t \in K$ we have $|x - \mathbb{E}(U(t))| < \delta$. So, applying (10) in the case where $H = \mathbb{R}$ we get $|x(t) - \mathbb{E}_C(U(t))| < \epsilon$, and taking the sumpremum over all $t \in K$ we obtain $\|x - \mathbb{E}_C(U)\|_\infty < \epsilon$. Therefore, the restriction of the identity operator on $C(K)$ to $B_{C(K)}$ is midpoint continuous at every extreme point of this set. Let us mention that if K is an infinite compact, the restriction of the identity operator on $C(K)$ to $B_{C(K)}$ is not slicely continuous at any element of the unit sphere of $C(K)$ (it is well known that every slice of $B_{C(K)}$ has diameter 2).

At this point, we state the counterpart of the above concept in terms of countable coverings. To avoid measurability problems, from now on we consider only simple random variables. Let us observe that if Φ is a map between the normed spaces X and Y and $U : \Omega \longrightarrow X$ is a simple random variable, then $\Phi \circ U$ is a simple random variable too.

DEFINITION 11. *Let X and Y be normed spaces, let $A \subseteq X$, and let $\epsilon > 0$. A map $\Phi : X \longrightarrow Y$ is said to be ϵ-σ-midpoint continuous on A if we can write*

$$A = \bigcup_{n \in \mathbb{N}} A_n$$

in such a way that for every $x \in A_n$ there is a number $\delta > 0$ such that if U is a simple random variable on the probability space (Ω, Σ, P) with values in A_n satisfying $\|x - \mathbb{E}(U)\| < \delta$, then for every measurable set $C \subset \Omega$ such that $P(C) \geq \frac{1}{2}$ we have $\|\Phi x - \mathbb{E}_C(\Phi U)\| \leq \epsilon$.

The map Φ is σ-midpoint continuous if it is ϵ-σ-midpoint continuous for each $\epsilon > 0$.

According to Theorem 3 and Lemma 9 it follows that if T is a bounded linear operator between two normed spaces X and Y, one of them being MLUR renormable, then T is σ-midpoint continuous on X. Also, from those results we have that the space X admits an equivalent MLUR norm if and only if the identity operator on X is σ-midpoint continuous.

As an example, we shall show the σ-midpoint continuity of the identity operator when X is a Hilbert space using probabilistic tools. Let $\epsilon > 0$, and define, for each positive rational number r, the set

$$X_r = \left(r + \frac{\epsilon^2}{1 + r^2} \right) B_X \setminus r B_X.$$

The sequence $\{X_r\}_{r \in \mathbb{Q}^+}$ is clearly a covering of $X \setminus \{0\}$. Let $x \in X \setminus \{0\}$, and choose $r \in \mathbb{Q}^+$ such that $x \in X_r$. Proceeding as in (9) we can deduce that if U is a (simple) random variable with values in X_r such that $\|x - \mathbb{E}(U)\| < \epsilon^2 \|x\|/(1 + r^2)$, then $(\mathbb{E}(\|x - U\|))^2 < 8\,\epsilon^2$. It follows that the identity operator on X is σ-midpoint continuous (the same argument shows that this operator is σ-slicely continuous). Let us mention that the construction of the above countable covering is based on the methods used in [12] and [13] to characterize the classes of normed spaces that admit an equivalent LUR or MLUR renorming.

The next result provides an example of σ-midpoint continuous map in the particular setting of spaces of bounded functions.

LEMMA 12. *Let V be a set, let $\epsilon > 0$, and let $V_1, \ldots, V_m$ be subsets of V. Assume that for each $r \in \{1, \ldots, m\}$ the identity map on $\ell_\infty(V_r)$ is ϵ-σ-midpoint continuous on some set $A_r \subseteq \ell_\infty(V_r)$. Then the map $\Phi : \ell_\infty(V) \to \ell_\infty(V)$ defined by the formula*

$$\Phi x = x \cdot \mathbb{1}_{\bigcup_{r=1}^m V_r}$$

is ϵ-σ-midpoint continuous on the set $\{x \in \ell_\infty(V) : x_{\restriction V_r} \in A_r \text{ for all } r = 1, \ldots, m\}$. (The symbol $x_{\restriction V_r}$ denotes the restriction of the function x to the set V_r.)

Proof. By hypothesis, for each $r = 1, \ldots, m$ we can write

$$A_r = \bigcup_{n \in \mathbb{N}} A_{r,n}$$

in such a way that every $z \in A_{r,n}$ is an (ϵ, δ)-strongly extreme point of $\mathrm{co}(A_{r,n})$, for some $\delta = \delta(r, n, z) > 0$.

For each $s = (s_1, \ldots, s_m) \in \mathbb{N}^m$ we define the set

$$A_s = \{x \in A : x_{\restriction V_r} \in A_{r,s_r} \quad \forall r = 1, \ldots, m\}.$$

It is clear that $A = \bigcup_{s \in \mathbb{N}^m} A_s$. Let us fix $s = (s_1, \ldots, s_m) \in \mathbb{N}$ and $x \in A_s$. Let $U : (\Omega, \Sigma, P) \to A_s$ be a simple random variable such that $\|x - \mathbb{E}(U)\|_\infty < \delta$, where $\delta = \min\{\delta(r, s_r, x) : r = 1, \ldots, m\}$. Let, for each $r = 1, \ldots, m$, U_r be the random variable defined by the formula $U_r(w) = U(w)_{\restriction V_r}$, $w \in \Omega$. Then U_r takes its values

in A_{r,s_r} and $\|x_{\restriction V_r} - \mathbb{E}(U_r)\|_\infty < \delta$. Since $x_{\restriction V_r}$ is an (ϵ,δ)-strongly extreme point of $\mathrm{co}(A_{r,s_r})$, from Lemma 9 it follows that if C is a measurable subset of Ω with $P(C) \geq \frac{1}{2}$, then $\|x_{\restriction V_r} - \mathbb{E}_C(U_r)\|_\infty \leq \epsilon$. As this inequality holds for each $r = 1, \ldots, m$ we get

$$\|\Phi x - \mathbb{E}_C(\Phi U)\|_\infty = \max_{r=1,\ldots,m} \|x_{\restriction V_r} - \mathbb{E}_C(U_r)\|_\infty \leq \epsilon,$$

as we wanted to show. □

Remark. In [7, Lemma 5.2] (see also [13, Lemma 3]) the author considered, for any set V and any $\epsilon > 0$, the set $E_\epsilon(V)$ made up of all the functions $x \in \ell_\infty(V)$ for which there exist $a, b \in \mathbb{R}$ and a binary partition $\{M, N\}$ of V such that $\|x - (a\mathbb{1}_M + b\mathbb{1}_N)\|_\infty < \epsilon$ and showed that the identity map of $\ell_\infty(V)$ is 15 ϵ-midpoint continuous on the set $E_\epsilon(V)$.

As may be conjectured, the class of σ-midpoint continuous maps includes that of the σ-slicely continuous maps.

PROPOSITION 13. *Let X and Y be normed spaces, let $A \subseteq X$, and let $\Phi : A \longrightarrow Y$ be a map. If Φ is σ-slicely continuous, then Φ is σ-midpoint continuous on A.*

Proof. Let us fix $\epsilon > 0$. According to Theorem 5 we can write

$$A = \bigcup_{n\in\mathbb{N}} A_n$$

in such a way that for every $x \in A_n$ there exists a number $\delta > 0$ with the property that $\mathbb{E}(\|\Phi x - \Phi U\|) < \epsilon$ whenever $U : (\Omega, \Sigma, P) \to A_n$ is a (simple) random variable such that $\|x - \mathbb{E}(U)\| < \delta$. So, for such a vector x, such a random variable U, and every measurable set $C \subset \Omega$ with $P(C) \geq \frac{1}{2}$ we have

$$\|\Phi x - \mathbb{E}_C(\Phi U)\| \leq \frac{1}{P(C)}\,\mathbb{E}(\|x - \Phi U\|) \leq 2\,\mathbb{E}(\|x - \Phi U\|) < 2\,\epsilon. \qquad \square$$

Note that the converse of this result does not hold in general. If X is MLUR renormable space with no equivalent LUR renorming, then the identity operator $Id : X \to X$ is σ-midpoint continuous but not σ-slicely continuous.

In [14], some stability properties for the class of σ-slicely continuous maps are established. There, it is shown that linear combinations, norm pointwise limits, and composition of σ-slicely continuous maps (as well as the composition of a continuous map $B : Y \longrightarrow Z$ with a σ-slicely continuous map $\Psi : X \longrightarrow Y$) are σ-slicely continuous too. The following results show that some of these properties are also shared by the class of σ-midpoint continuous maps.

PROPOSITION 14. *Let X and Y be normed spaces, let $\alpha, \beta \in \mathbb{R}$, and let Φ_1 and Φ_2 be two maps from X into Y, which are ϵ-σ-midpoint continuous on a set $A \subseteq X$. Then the map $\Phi = \alpha\Phi_1 + \beta\Phi_2$ is $(|\alpha| + |\beta|)\,\epsilon$-$\sigma$-midpoint continuous on A.*

In particular, linear combinations of σ-midpoint continuous maps are σ-midpoint continuous too.

Proof. It is straightforward to show that if $\{A_{1,n}\}_n$ and $\{A_{2,n}\}_n$ are coverings of A satisfying the conditions of Definition 11 for the maps Φ_1 and Φ_2, respectively, then the sequence of sets $\{A_{1,n} \cap A_{2,m} : n, m \in \mathbb{N}\}$ constitutes a covering of this type for the map Φ. □

PROPOSITION 15. *Let X and Y be normed spaces, let $\epsilon > 0$, and let Φ be a map from X into Y. Assume that there is a sequence $\{\Phi_n\}_n$ of maps from X into Y such that for all $x \in X$ there exists $m = m(x) \in \mathbb{N}$ satisfying*

$$\|\Phi x - \Phi_m x\| < \epsilon.$$

If for each $n \in \mathbb{N}$ the map Φ_n is ϵ-σ-midpoint continuous on X, then the map Φ is $3\,\epsilon$-σ-midpoint continuous on X.

In particular, the norm pointwise limit of a sequence of σ-midpoint continuous maps is also σ-midpoint continuous.

Proof. Let us define, for each $m \in \mathbb{N}$, the set

$$X_m = \{x \in X : \|\Phi x - \Phi_m x\| < \epsilon\}.$$

By the hypothesis we have $X = \bigcup_{m\in\mathbb{N}} X_m$.

Let $m \in \mathbb{N}$. Since the map Φ_m is ϵ-σ-midpoint continuous on X_m, we can write $X_m = \bigcup_{n\in\mathbb{N}} X_{m,n}$ in such a way that for every $x \in X_{m,n}$ there is a number $\delta = \delta_x > 0$ with the property that if $U : (\Omega, \Sigma, P) \to X_{m,n}$ is a simple random variable, and $\|x - \mathbb{E}(U)\| < \delta_x$, then for every measurable set $C \subseteq \Omega$ such that $P(C) \geq \frac{1}{2}$ we have $\|\Phi_m x - \mathbb{E}_C(\Phi_m U)\| \leq \epsilon$.

Hence, for such a point x, such a set C, and such a simple random variable U we have

$$\begin{aligned} \|\Phi x - \mathbb{E}_C(\Phi U)\| &\leq \|\Phi x - \Phi_m x\| + \|\Phi_m x - \mathbb{E}_C(\Phi_m U)\| \\ &\quad + \|\mathbb{E}_C(\Phi_m U - \Phi U)\| \\ &< \epsilon + \|\Phi x - \Phi_m x\| + \|\mathbb{E}_C(\Phi_m U - \Phi U)\|. \end{aligned} \tag{11}$$

On the other hand, by the definition of X_m we have $\|\Phi_m U(w) - \Phi U(w)\| < \epsilon$, for each $w \in \Omega$, and $\|\Phi x - \Phi_m x\| < \epsilon$. Combining these inequalities with (11) we obtain

$$\|\Phi x - \mathbb{E}_C(\Phi U)\| < 3\epsilon,$$

and the proposition is proved. □

Remark. A variant of the proof of Theorem 5 shows that if $\Phi : X \longrightarrow Y$ and $\Psi : Y \longrightarrow Z$ are, respectively, a σ-midpoint continuous map and a σ-slicely continuous map, then the map $\Xi = \Psi \circ \Phi$ is σ-midpoint continuous. It is also easy to show that the map Ξ is σ-midpoint continuous whenever Ψ is a bounded linear operator and Φ is σ-midpoint continuous, or whenever Φ and Ψ are bounded linear operators and one of them is σ-midpoint continuous. We do not know whether the composition of two σ-midpoint continuous maps is σ-midpoint continuous too. In the next section, we will see that the composition of a continuous and a σ-midpoint continuous map is not necessarily σ-midpoint continuous.

3. Renorming results. In this section we develop various techniques for constructing σ-midpoint continuous maps. As an application of these techniques, we obtain some results about MLUR renormings of Banach spaces.

In [10] it is shown that a normed space X is MLUR renormable whenever there exist a σ-slicely continuous map $\Psi : X \longrightarrow X$ and an MLUR renormable subspace $Y \subset X$ such that $x - \Psi x \in Y$ for all $x \in X$. The use of σ-midpoint continuous maps enables us to give the following improvement of that result.

THEOREM 16. *Let X and Y be normed spaces. Assume that there exist an MLUR renormable subspace $Y_1 \subset Y$, a bounded linear map $T : X \longrightarrow Y$, and a σ-midpoint continuous map $\Psi : X \longrightarrow Y$ such that the map $\Phi = T + \Psi$ takes its values in Y_1. Then the maps T and Φ are σ-midpoint continuous too.*

Proof. By Proposition 14 it is enough to show that T is σ-midpoint continuous. Let us fix $\epsilon > 0$. Since the space Y_1 is MLUR renormable we can write

$$Y_1 = \bigcup_{n,k\in\mathbb{N}} Y_{n,k}$$

in such a way that every $y \in Y_{n,k}$ is an $\left(\frac{\epsilon}{2}, \frac{1}{k}\right)$-strongly extreme point of $\mathrm{co}(Y_{n,k})$.

For each $n, k \in \mathbb{N}$ with $k > \frac{2}{\epsilon}$ we set $X_{n,k} = \Phi^{-1}(Y_{n,k})$.

As the map Ψ is σ-midpoint continuous on X there is a countable decomposition of $X_{n,k}$,

$$X_{n,k} = \bigcup_{p \in \mathbb{N}} X_{n,k,p},$$

in such a way that for each $x \in X_{n,k,p}$ there exists a number $0 < \delta_x < \frac{1}{2\|T\|k}$ with the property that if $U : (\Omega, \Sigma, P) \to X_{n,k,p}$ is a simple random variable and $\|x - \mathbb{E}(U)\| < \delta_x$, then for every measurable set $C \subset \Omega$ such that $P(C) \geq \frac{1}{2}$ we have

$$\|\Psi x - \mathbb{E}_C(\Psi U)\| \leq \frac{1}{2k}. \tag{12}$$

It is clear that the family of sets $\left\{X_{n,k,p} : n, k, p \in \mathbb{N},\ k > \frac{2}{\epsilon}\right\}$ is a countable covering of X.

Fix $n, k, p \in \mathbb{N}$ and $x \in X_{n,k,p}$, and let U be a simple random variable on the probability space (Ω, Σ, P) with values in $X_{n,k,p}$ such that $\|x - \mathbb{E}(U)\| < \delta_x$. Then

$$\begin{aligned}\|\Phi x - \mathbb{E}(\Phi U)\| &= \|\mathbb{E}(Tx - TU) + \Psi x - \mathbb{E}(\Psi U)\| \\ &\leq \|T\|\|x - \mathbb{E}(U)\| + \|\Psi x - \mathbb{E}(\Psi U)\| \\ &< \frac{1}{2k} + \|\Psi x - \mathbb{E}(\Psi U)\|.\end{aligned}$$

From this inequality and (12) (with $C = \Omega$) we get $\|\Phi x - \mathbb{E}(\Phi U)\| < \frac{1}{k}$. Since ΦU is a simple random variable with values in $Y_{n,k}$ and Φx is an $\left(\frac{\epsilon}{2}, \frac{1}{k}\right)$-strongly extreme point of $\mathrm{co}(Y_{n,k})$, from Lemma 9 it follows that if C is a measurable subset of Ω such that $P(C) \geq \frac{1}{2}$, then

$$\|\Phi x - \mathbb{E}_C(\Phi U)\| < \frac{\epsilon}{2}.$$

This inequality together with (12) implies that

$$\|Tx - \mathbb{E}_C(TU)\| \leq \|\Phi x - \mathbb{E}_C(\Phi U)\| + \|\Psi x - \mathbb{E}_C(\Psi U)\| < \frac{\epsilon}{2} + \frac{1}{2k} < \epsilon.$$

So, the map T is σ-midpoint continuous on X, as we wanted to show. □

As a consequence of this theorem we obtain the following result of G. Alexandrov concerning the three space problem for the MLUR property.

Corollary 17 (Alexandrov [1]). *Let X be a Banach space. Suppose that there is a closed MLUR renormable subspace $Y \subset X$ such that the quotient X/Y has an equivalent LUR norm. Then X admits an equivalent MLUR norm.*

Proof. Let Q denote the quotient map from X onto X/Y. According to the Bartle–Graves theorem (see, e.g., [5, Chapter VII.3]), there is a continuous map $B : X/Y \longrightarrow X$ such that $BQx \in Qx$ for all $x \in X$. If we define $\Phi = BQ$, then $x - \Phi x \in Y$ for every $x \in X$. Since the space X/Y is LUR renormable the maps Q and Φ are σ-slicely continuous. Therefore, Φ is σ-midpoint continuous. Applying Theorem 16 we deduce that the identity operator on X is σ-midpoint continuous. □

Let us mention that the property of having an equivalent MLUR norm is not a three space property. In [7] the existence of a Banach space X and a closed subspace

$Y \subset X$ is shown such that Y and X/Y both have an equivalent MLUR norm while X does not. This counterexample reveals that the composition of a continuous map and a σ-midpoint continuous map is not necessarily σ-midpoint continuous. Indeed, let Q be the quotient map from X onto X/Y and B be a Bartle–Graves continuous selector of Q^{-1}. Since Q is linear and bounded and the space X/Y is MLUR renormable it follows that Q is σ-midpoint continuous on X. Nevertheless, the map $\Phi = B \circ Q$ does not have this property (otherwise, the identity operator of X should be σ-midpoint continuous by Theorem 16).

The following theorem provides another approach for constructing σ-midpoint continuous maps.

THEOREM 18. *Let X and Y be normed spaces, let $\epsilon > 0$, and let Λ be a set. Let $\tau : X \longrightarrow \Lambda$ be a σ-slicely constant map, and let, for each $\alpha \in \Lambda$, $\psi_\alpha : X \longrightarrow Y$ be a map which is ϵ-σ-midpoint on some set $X_\alpha \subseteq X$. Then the map $\Psi : X \longrightarrow Y$ defined by the formula*

$$\Psi x = \psi_{\tau(x)}(x)$$

is $2\,\epsilon$-σ-midpoint continuous on the set $A = \{x \in X : x \in X_{\tau(x)}\}$.

Proof. For each $\alpha \in \Lambda$ we have a countable covering of X_α,

$$X_\alpha = \bigcup_{n \in \mathbb{N}} X_{\alpha,n},$$

in such a way that for every $x \in X_{\alpha,n}$ there is a number $0 < \delta_{\alpha,n,x} < \epsilon$ with the property that if $U : (\Omega, \Sigma, P) \to X_{\alpha,n}$ is a simple random variable such that $\|x - \mathbb{E}(U)\| < \delta_{\alpha,n,x}$, then for any measurable subset $C \subset \Omega$ with $P(C) \geq \frac{1}{2}$ we have

$$\|\psi_\alpha x - \mathbb{E}_C(\psi_\alpha U)\| < \epsilon. \tag{13}$$

Let $n, k \in \mathbb{N}$, and let $A_{n,k}$ denote the set made up of all the vectors $x \in X_{\tau(x),n,k}$ such that $\|x\| \leq k$, $\|\Psi x\| \leq k$, and $\delta_{\tau(x),n,x} > \frac{1}{k}$. It is clear that

$$A = \bigcup_{n,k \in \mathbb{N}} A_{n,k}.$$

Since the map τ is σ-slicely constant, for each $n, k \in \mathbb{N}$ we get another decomposition of $A_{n,k}$,

$$A_{n,k} = \bigcup_{m \in \mathbb{N}} A_{n,k,m},$$

with the property that for every $x \in A_{n,k,m}$ there is an open half space $H_x \subset X$ such that

$$\tau(u) = \tau(x) \quad \text{whenever} \quad u \in H_x \cap A_{n,k,m}. \tag{14}$$

Applying Lemma 6 to the set $A_{n,k,m}$, with $\xi = \frac{1}{4k^2}$, we can write

$$A_{n,k,m} = \bigcup_{p \in \mathbb{N}} A_{n,k,m,p}$$

in such a way that for every $x \in A_{n,k,m,p}$ there exist a number $\eta = \eta_x > 0$, an open half space $H'_x \subset X$, and an element $w_x \in A_{n,k,m}$ such that

$$x \in H'_x \cap A_{n,k,m,p} \subseteq H_{w_x} \cap A_{n,k,m} \tag{15}$$

and if $U : (\Omega, \Sigma, P) \to A_{n,k,m,p}$ is a random variable with $\|x - \mathbb{E}(U)\| < \eta_x$, and $N = \{w \in \Omega : U(w) \notin H'_x\}$, then

$$P(N) < \frac{1}{4k^2}. \tag{16}$$

Now let us fix natural numbers n, k, m, and p such that $A_{n,k,m,p} \neq \emptyset$. Take $x \in A_{n,k,m,p}$, and let $U : (\Omega, \Sigma, P) \to A_{n,k,m,p}$ be a simple random variable satisfying $\|x - \mathbb{E}(U)\| < \min\{\eta_x, \frac{1}{2k}\}$. Our goal is to show that if C is a measurable subset of Ω such that $P(C) \geq \frac{1}{2}$, then

$$\|\Psi x - \mathbb{E}_C(\Psi U)\| < 2\,\epsilon. \tag{17}$$

We will prove first that

$$\|x - \mathbb{E}(U)\| < \frac{1}{k} \quad \text{and} \quad U(w) \in H'_x \quad \forall w \in \Omega \Rightarrow \|\Psi x - \mathbb{E}_C(\Psi U)\| < \epsilon. \tag{18}$$

Thanks to (15) it follows that $U(w) \in A_{n,k,m} \cap H_{w_x}$ for all $w \in \Omega$. Using (14) we deduce that

$$\tau(U(w)) = \tau(x). \tag{19}$$

Hence, $U(w) \in X_{\tau(x),n,k}$. As $\delta_{\tau(x),n,x} > \frac{1}{k}$, thanks to (13) it follows that

$$\left\|\psi_{\tau(x)} x - \mathbb{E}_C(\psi_{\tau(x)} U)\right\| < \epsilon.$$

From (19) we also have $\psi_{\tau(x)} = \psi_{\tau(U(w))}$. Therefore,

$$\left\|\psi_{\tau(x)} x - \mathbb{E}_C(\psi_{\tau(U(w))} U(w))\right\| < \epsilon$$

and assertion (18) is proved.

To finish, suppose that not all the vectors $U(w)$ necessarily belong to the half space H'_x. Let $N = \{w \in \Omega : U(w) \notin H'_x\}$, and let U' be the simple random variable defined by the formula $U' = x\mathbb{1}_N + U\mathbb{1}_{\Omega\setminus N}$.

From inequality (16) and the fact $A_{n,k,m,p} \subseteq kB_X$ we get

$$\|x - \mathbb{E}(U')\| \leq \|\mathbb{E}(x - U)\| + P(N)\|\mathbb{E}_N(x - U)\| < \frac{1}{2k} + 2k\frac{1}{4k^2} = \frac{1}{k}.$$

Taking into account that U' takes its values in $A_{n,k,m,p} \cap H'_x$, from (18) it follows that $\|\Psi x - \mathbb{E}_C(\Psi U')\| < \epsilon$. Using again inequality (16) and the fact $\Psi(A_{n,k,m,p}) \subseteq kB_X$ we obtain

$$\begin{aligned}\|\Psi x - \mathbb{E}_C(\Psi U)\| &\leq \|\Psi x - \mathbb{E}_C(\Psi U')\| + \|\mathbb{E}_C(\Psi U - \Psi U')\| \\ &\leq \epsilon + P(N)\|\mathbb{E}_{C\cap N}(\Psi x - \Psi U)\| < 2\,\epsilon,\end{aligned}$$

as we wanted. □

Now we apply Theorem 18 and Proposition 8 to get some results about MLUR renormability in those Banach spaces X for which there exist a set Γ and a σ-slicely continuous map $\Phi : X \longrightarrow c_0(\Gamma)$.

COROLLARY 19. *Let X be a normed space, and let $\Phi : X \longrightarrow c_0(\Gamma)$ be a σ-slicely continuous map. Suppose that there exists a family $\{T_\gamma\}_{\gamma\in\Gamma}$ of bounded linear operators on X with the following properties:*

(1) *for each* $\gamma \in \Gamma$, $T_\gamma X$ *is an MLUR renormable subspace of* X*;*
(2) *for each* $x \in X$,

$$x \in \overline{\operatorname{span}\{T_\gamma x : \gamma \in \operatorname{supp}\Phi x\}}^{\|\cdot\|}.$$

Then X *admits an equivalent MLUR norm.*

Proof. Let us fix $\epsilon > 0$. We are going to demonstrate the existence of a sequence $\{\Psi_n\}_n$ of ϵ-σ-midpoint continuous maps on X such that

$$\lim_n \|x - \Psi_n x\| = 0$$

for all $x \in X$.

For technical reasons we introduce on Γ a well order $\prec$. Let Λ be the family of all finite sets $\{(\gamma_1, r_1), \dots, (\gamma_m, r_m)\} \in 2^{\Gamma\times\mathbb{Q}}$ such that $\gamma_i \neq \gamma_j$ whenever $i \neq j$, and let us define, for each $\alpha = \{(\gamma_1, r_1), \dots, (\gamma_m, r_m)\} \in \Lambda$, the map

$$\psi_\alpha = \sum_{i=1}^{m} r_i T_{\gamma_i}.$$

Since for each $i = 1, \dots, m$ the range of T_{γ_i} is an MLUR renormable space, it follows from Proposition 14 that the map ψ_α is ϵ-σ-midpoint continuous on X.

Let $n \in \mathbb{N}$. For each $x \in X$ we can choose a finite set

$$\Delta(x, n) = \{\gamma_1(x, n) \prec \cdots \prec \gamma_j(x, n)\} \subset \operatorname{supp}\Phi x$$

and a vector $(r_1(x, n), \dots, r_j(x, n)) \in \mathbb{Q}^j \setminus \{0\}$ such that

$$\left\| x - \sum_{i=1}^{j} r_i(x, n) T_{\gamma_i(x,n)} x \right\| < 1/n. \tag{20}$$

Now we define a map $\tau_n : X \longrightarrow \Lambda$ by setting, for each $x \in X$,

$$\tau_n(x) = \{(\gamma_i(x, n), r_i(x, n))\}_{i=1}^{j}.$$

This map is σ-slicely constant. Indeed, let $m \in \mathbb{N}$, $q \in \mathbb{Q}\cap(0, 1)$, and $r = (r_1, \dots, r_m) \in \mathbb{Q}^m$. Let $X_{m,q,r}$ denote the set made up of the vectors $x \in X$ such that $\Delta(x, n) \subseteq M_{\Phi,q}(x)$, $\#M_{\Phi,q}(x) = m$, and if $M_{\Phi,q}(x) = \{\gamma_1 \prec \cdots \prec \gamma_m\}$ and $\Delta(x, n) = \{\gamma_{t_1} \prec \cdots \prec \gamma_{t_j}\}$, then

$$r_i = 0 \quad \text{if and only if} \quad \gamma_i \notin \Delta(x, n) \tag{21}$$

and

$$r_i(x, n) = r_{t_i} \quad \text{for } i = 1, \dots, j. \tag{22}$$

It is clear that

$$X = \bigcup_{m\in\mathbb{N},\, q\in\mathbb{Q}\cap(0,1),\, r\in\mathbb{Q}^m} X_{m,p,r}.$$

Fix $m \in \mathbb{N}$, $q \in \mathbb{Q} \cap (0, 1)$, and $r \in \mathbb{Q}^m$. By Proposition 8 we can write

$$X_{m,q,r} = \bigcup_{p\in\mathbb{N}} X_{m,q,r,p}$$

in such a way that for every $p \in \mathbb{N}$ and every $x \in X_{m,q,r,p}$ there exists an open half space $H_x \subset X$ such that $x \in H_x$ and $M_{\Phi,q}(u) = M_{\Phi,q}(x)$ whenever $u \in H_x \cap X_{m,q,r,p}$. Combining this equality with conditions (21) and (22) we deduce that $\Delta(u,n) = \Delta(x,n)$ and that $r_i(u,n) = r_i(x,n)$ for each $i = 1,\dots,m$. In particular $\tau_n(u) = \tau_n(x)$, and the map τ_n is σ-slicely constant.

Now for each $x \in X$ we define

$$\Psi_n x = \psi_{\tau_n(x)} x = \sum_{i=1}^{j} r_i(x,n) T_{\gamma_i(x,n)} x.$$

According to Theorem 18 it follows that the map Ψ_n is ϵ-σ-midpoint continuous on X.

On the other hand, from (20) we get $\lim_n \|x - \lim_n \Psi_n x\| = 0$ for each $x \in X$. Applying Proposition 15 we deduce that the identity operator on X is $3\,\epsilon$-σ-midpoint continuous. As this assertion is true for every $\epsilon > 0$, the space X has an equivalent MLUR norm. □

As a consequence of this result, we obtain the following corollary, which is an MLUR version of the classical Zizler criterion in [19] about LUR renormability in Banach spaces with projectional resolutions of the identity.

COROLLARY 20. *Let X be a normed space, and let $\{T_\gamma\}_{\gamma\in\Gamma}$ be a family of bounded linear operators on X with the following properties:*

(1) *for every $\gamma \in \Gamma$, $T_\gamma X$ is an MLUR renormable subspace of X;*
(2) *for every $x \in X$, $\{\|T_\gamma x\|\}_{\gamma\in\Gamma} \in c_0(\Gamma)$, and*

$$x \in \overline{\operatorname{span}\{T_\gamma x : \gamma \in \Gamma\}}^{\|\cdot\|}.$$

Then X admits an equivalent MLUR norm.

Proof. It is enough to show that the map $\Phi : X \longrightarrow c_0(\Gamma)$ defined by the formula

$$\Phi x = \{\|T_\gamma x\|\}_{\gamma\in\Gamma}$$

is σ-slicely continuous. For all $x, y \in X$ and $\lambda, \mu \geq 0$ such that $\lambda + \mu = 1$ we have

$$0 \leq \Phi(\lambda x + \mu y) \leq \lambda \Phi x + \mu \Phi y,$$

and the result follows from the fact that every positive convex map with values in an LUR lattice is σ-slicely continuous. □

Our last application of Theorem 18 is the following corollary, which is a generalization of the main tool used by Haydon in [7] for the construction of MLUR norms in spaces of continuous functions on trees.

COROLLARY 21. *Let K be a locally compact space. Assume that there exist a σ-slicely continuous map $\Phi : C_0(K) \longrightarrow c_0(\Gamma)$ and a family $\{K_\gamma\}_{\gamma\in\Gamma}$ of closed and open subsets of K with the following properties:*

(1) *for each $\gamma \in \Gamma$ and each $\epsilon > 0$ there is a set $Y_{\gamma,\epsilon} \subseteq C(K_\gamma)$ such that the identity operator of $C_0(K_\gamma)$ is ϵ-σ-midpoint continuous on $Y_{\gamma,\epsilon}$;*
(2) *for each $\epsilon > 0$, each $x \in C_0(K)$, and $t \in K$ with $x(t) \neq 0$, there exists an element $\gamma \in \operatorname{supp} \Phi x$ such that $t \in K_\gamma$ and $x_{\restriction K_\gamma} \in Y_{\gamma,\epsilon}$.*

Then the space $C_0(K)$ is MLUR renormable.

Proof. Let $X = C_0(K)$. As in the proof of Corollary 19 we introduce a well order $\prec$ on the set Γ. Fix $\epsilon > 0$, and let Λ denote the set of all $\alpha \in 2^\Gamma$ such that $\#\alpha < \infty$.

Let $n \in \mathbb{N}$ such that $\epsilon > \frac{1}{n}$. Let us define, for each $\alpha \in \Lambda$,

$$\psi_\alpha x = x \cdot \mathbb{1}_{\bigcup_{\gamma\in\alpha} K_\gamma}.$$

It is clear that $\psi_\alpha x \in C_0(K)$ for all $x \in C_0(K)$, and from Lemma 12 it follows that the map ψ_α is ϵ-σ-midpoint continuous on the set

$$A_{\alpha,n} = \left\{ x \in C_0(K) : x_{\restriction K_\gamma} \in Y_{\gamma,\frac{1}{n}} \quad \forall \gamma \in \alpha \right\}.$$

On the other hand, because of the hypothesis, for each $x \in X$ we have

$$\{t \in K : |x(t)| \geq 1/n\} \subseteq \bigcup \left\{ K_\gamma : \gamma \in \operatorname{supp} \Phi x \text{ and } x_{\restriction K_\gamma} \in Y_{\gamma,\frac{1}{n}} \right\}.$$

By compactness we deduce the existence of a finite set $\Delta_n(x) \subseteq \operatorname{supp} \Phi x$ such that

$$\|x - x \cdot \mathbb{1}_{\bigcup_{\gamma \in \Delta_n(x)} K_\gamma}\|_\infty < 1/n \tag{23}$$

and

$$x_{\restriction K_\gamma} \in Y_{\gamma,1/n} \quad \forall \gamma \in \Delta_n(x). \tag{24}$$

Using the same argument as in Corollary 20 we get that the map $\tau_n : X \longrightarrow \Lambda$ defined by the formula

$$\tau_n(x) = \Delta_n(x), \quad x \in C_0(K),$$

is σ-slicely constant.

Now for each $x \in X$ we set

$$\Psi_n x = \psi_{\tau_n(x)} x = x \cdot \mathbb{1}_{\bigcup_{\gamma \in \Delta_n(x)} K_\gamma}.$$

According to Theorem 18 we deduce that Ψ_n is $2\,\epsilon$-σ-midpoint continuous on the set

$$A_n = \left\{ x \in C_0(K) : x \in X_{\tau(x)} \right\} = \left\{ x \in C_0(K) : x \in x_{\restriction K_\gamma} \in Y_{\gamma,\frac{1}{n}} \right\}.$$

By (24) we have $A_n = X$. Moreover, from (23) we get $\lim_n \|x - \Psi_n x\| = 0$ for each $x \in X$. Applying Proposition 15 we deduce that the identity operator on X is $6\,\epsilon$-σ-midpoint continuous. As this assertion holds for every $\epsilon > 0$, it follows that the space X is MLUR renormable. □

REFERENCES

[1] G. Alexandrov, *On the three space problem for MLUR renorming of Banach spaces*, C. R. Acad. Bulgare Sci., 42 (1989), pp. 17–20.

[2] G. Alexandrov and I. Dimitrov, *On equivalent weakly midpoint locally uniformly rotund renormings of the space ℓ_∞*, in Proceedings of the 14th Annual Spring Conference of the Union of Bulgarian Mathematicians, Sunny Beach, Bulgaria, 1985, pp. 189–191 (in Russian).

[3] E. Asplund and R. T. Rockafellar, *Gradients of convex functions*, Trans. Amer. Math. Soc., 139 (1969), pp. 443–467.

[4] I. Cioranescu, *Geometry of Banach Spaces, Duality Mappings and Nonlinear Problems*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1990.

[5] R. Deville, G. Godefroy, and V. Zizler, *Smoothness and Renormings in Banach Spaces*, Pitman Monogr. Surveys Pure Appl. Math. 64, Longman Scientific and Technical, Harlow, UK, 1993.

[6] J. Diestel and J. J. Uhl, *Vector Measures*, Math. Surveys 15, AMS, Providence, RI, 1977.

[7] R. Haydon, *Trees in renorming theory*, Proc. London Math. Soc. (3), 78 (1999), pp. 541–585.

[8] Z. Hu, W. B. Moors, and M. A. Smith, *On a Banach space without a weak mid-point locally uniformly rotund norm*, Bull. Austral. Math. Soc., 56 (1997), pp. 193–196.

[9] K. Kunen and H. Rosenthal, *Martingale proofs of some geometric results in Banach space theory*, Pacific J. Math., 100 (1982), pp. 153–175.
[10] S. Lajara and A. J. Pallarés, *A nonlinear map for midpoint locally uniformly rotund renorming*, Bull. Austral. Math. Soc., 72 (2005), pp. 39–44.
[11] J. Lindenstrauss, *Weakly compact sets—Their topological properties and Banach spaces they generate*, in Symposium on Infinite-Dimensional Topology, Ann. of Math. Stud. 69, Princeton University Press, Princeton, NJ, 1972, pp. 235–273.
[12] A. Moltó, J. Orihuela, and S. Troyanski, *Locally uniformly rotund renorming and fragmentability*, Proc. London Math. Soc. (3), 75 (1997), pp. 619–640.
[13] A. Moltó, J. Orihuela, S. Troyanski, and M. Valdivia, *Midpoint locally uniform rotundity and a decomposition method for renorming*, Q. J. Math., 52 (2001), pp. 181–193.
[14] A. Moltó, J. Orihuela, S. Troyanski, and M. Valdivia, *A Nonlinear Transfer Technique for Renorming*, Lecture Notes in Math., Springer, to appear.
[15] A. Moltó, J. Orihuela, S. Troyanski, and M. Valdivia, *Continuity properties up to countable partition*, RACSAM Rev. R. Acad. Cienc. Exactas Fís. Nat. Ser. A Mat., 100 (2006), pp. 279–294.
[16] M. Raja, *On locally uniformly rotund norms*, Mathematika, 46 (1999), pp. 343–358.
[17] E. Zeidler, *Nonlinear Functional Analysis and Its Applications,* II/B, *Nonlinear Monotone Operators*, Springer, New York, 1990.
[18] E. Zeidler, *Nonlinear Functional Analysis and Its Applications,* III, *Variational Methods and Optimization*, Springer, New York, 1985.
[19] V. Zizler, *Locally uniformly rotund renorming and decomposition of Banach spaces*, Bull. Austral. Math. Soc., 29 (1984), pp. 259–265.

SIAM J. OPTIM.
Vol. 18, No. 3, pp. 1046–1060

SOLUTION CONTINUITY IN MONOTONE AFFINE VARIATIONAL INEQUALITIES*

STEPHEN M. ROBINSON†

Abstract. In this paper we study the behavior of solutions of finite-dimensional monotone affine variational inequalities posed over graph-convex polyhedral multifunctions. We identify precisely the class of positive semidefinite linear transformations appearing in these variational inequalities for which the solution sets will be Lipschitzian in the argument of the underlying multifunction. This class is that of *cocoercive* linear transformations, which include, but are not limited to, those appearing in problems of linear or of convex quadratic programming.

Key words. multifunctions, affine variational inequalities, monotone operators, Lipschitz continuity, polyhedral multifunctions, cocoercivity, Dunn property, psd-plus

AMS subject classifications. 49J53, 49J40, 49K40, 90C31, 90C33

DOI. 10.1137/060658576

1. Introduction.

In this paper we study the behavior of solutions of monotone affine variational inequalities in $\mathbb{R}^n$ posed over graph-convex polyhedral multifunctions. The goal is to identify precisely the class of positive semidefinite linear transformations appearing in these variational inequalities for which the solution sets will be Lipschitzian in the argument of the underlying multifunction. For the simplest type of constraining multifunction, this simply means that we are varying the right-hand sides of the linear equations and inequalities defining the set, and we want to be sure that the set of solutions of the variational inequality is a Lipschitzian multifunction of those right-hand sides. Our main results are that if no restrictions are placed on the underlying set, then this Lipschitzian behavior occurs always when the linear transformation has a property that has been called *cocoercivity* in the literature and never otherwise. Thus, the cocoercive linear transformations are the largest class for which this stability property holds. Such results were previously known for linear and for convex quadratic programming problems, both of which belong to the cocoercive class but do not exhaust it.

This paper has five sections, of which this is the first. In the remainder of this section we fix notation and discuss some of the terminology used below. Section 2 develops various properties relating polyhedrality and the Lipschitz condition; these are the underlying tools that we will use in the remainder of this paper. Section 3 studies monotone affine variational inequalities, developing properties of their solution sets under assumptions starting with positive semidefiniteness and then defining and characterizing the cocoercivity property. That section also briefly reviews existing

*Received by the editors April 30, 2006; accepted for publication (in revised form) June 5, 2007; published electronically October 4, 2007. The research reported here was sponsored in part by the National Science Foundation under grant DMS-0305930, in part by the Air Force Research Laboratory under agreement numbers FA9550-04-1-0192 and FA9550-07-1-0389, and in part by the U. S. Army Research Office under grant DAAD19-01-1-0502. The U. S. Government has certain rights in this material, and is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the author and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the sponsoring agencies or the U. S. Government.

http://www.siam.org/journals/siopt/18-3/65857.html

†Department of Industrial Engineering, University of Wisconsin–Madison, 1513 University Avenue, Madison, WI 53706-1539 (smrobins@wisc.edu).

literature on cocoercive linear transformations. Then, in section 4, we present the main results by first showing that cocoercivity plus a solvability condition suffices for Lipschitzian behavior, and then demonstrating that without cocoercivity one can define the underlying set in such a way that solutions of the problem fail even to be inner semicontinuous, much less Lipschitzian. Section 5 is an appendix containing the proof of one of the subsidiary results.

1.1. Notation and preliminaries. We give here some definitions required in the rest of the paper. These use extensively the relative topology induced on a subset X of $\mathbb{R}^n$ by the standard topology of $\mathbb{R}^n$. We will use the Euclidean norm throughout the paper.

The first definition defines a property originally called *upper Lipschitz continuity* [15, p. 208]. We have changed the term here to *outer Lipschitz continuity,* abbreviated OLC, in order to maintain some consistency with the terminology of [16]. The property is called *calmness* in that work [16, p. 399].

DEFINITION 1.1. *Let X be a subset of $\mathbb{R}^n$, x be a point of X, and $S : \mathbb{R}^n \to \mathbb{R}^m$ be a multifunction having closed values. S is* outer Lipschitz continuous at x relative to X with modulus λ *if there is some neighborhood V of x relative to X such that for each $x' \in V$ one has $S(x') \subset S(x) + \lambda\|x' - x\|B$, where B is the unit ball in $\mathbb{R}^m$.*

If F is OLC relative to X at a point $x \in X$, then it is also OLC relative to any subset of X that contains x. When X is not explicitly mentioned we take it to be the underlying space, in this case $\mathbb{R}^n$.

The second property of multifunctions that we require is inner semicontinuity. We will use the following definition rather than [16, Definition 5.4], but [16, Exercise 5.6] shows that the two are equivalent.

DEFINITION 1.2. *Let S be a multifunction from $\mathbb{R}^n$ to $\mathbb{R}^m$, X a subset of $\mathbb{R}^n$, and x a point of X. S is* inner semicontinuous at x relative to X *if for each open set Q that meets $S(x)$ there is a neighborhood V of x relative to X such that for each $x' \in V$, Q meets $S(x')$.*

For our purposes the most convenient measure of distance between sets will be the Pompeiu–Hausdorff distance. The following definition is compatible with [16, Example 4.13] if one adds the restriction that the sets U and V be nonempty and closed. We denote the closed ball of radius ρ about a point x of $\mathbb{R}^k$ by $B(x, \rho)$, and also use B to denote the unit ball $B(0, 1)$, the space involved being clear from the context.

DEFINITION 1.3. *The* Pompeiu–Hausdorff distance *between two subsets U and V of $\mathbb{R}^m$ is*

$$\rho[U, V] = \inf\{\eta \geq 0 \mid U \subset V + \eta B,\ V \subset U + \eta B\}.$$

Note that this distance may take the value $+\infty$.

The next definition applies the Pompeiu–Hausdorff distance to define Lipschitz continuity for multifunctions.

DEFINITION 1.4 (see [16, Def. 9.26]). *Let S be a multifunction from $\mathbb{R}^n$ to $\mathbb{R}^m$ having closed values on some subset X of $\mathbb{R}^n$, and let $\lambda \in \mathbb{R}_+$. S is* Lipschitz continuous *(equivalently,* Lipschitzian*) relative to X* with modulus λ *if for each x and x' in X, $\rho[S(x'), S(x)] \leq \lambda\|x' - x\|$.*

Note that to say S is Lipschitzian relative to X (or, as we often say below, on X) implies that a uniform modulus λ exists that works for each pair of points taken from X, even if that modulus is not explicitly mentioned. It is also often convenient

to speak of *local Lipschitz continuity* at a point $x \in X$, in which case the modulus need only exist for points in some neighborhood of x.

The following theorem is a slight sharpening of a result of Li [9, Theorem 2.1]. The sharpening consists of reducing the requirement of Hausdorff lower semicontinuity to inner semicontinuity. The proof, whose structure is similar to that of Li, is in section 5 (the appendix).

THEOREM 1.5. *Let S be a multifunction from $\mathbb{R}^n$ to $\mathbb{R}^m$ having closed values, let X be a convex subset of* $\operatorname{dom} S$*, and let λ be a nonnegative real number. The following are then equivalent:*

(a) *At each point of X, S is outer Lipschitz continuous relative to X with modulus λ and is inner semicontinuous relative to X.*

(b) *S is Lipschitz continuous relative to X with modulus λ.*

In the following sections we specialize the discussion first to polyhedral multifunctions and then to affine variational inequalities over polyhedral convex sets that need not be fixed, but may vary in a restricted way; specifically, they are sections of a graph-convex polyhedral multifunction.

2. Polyhedrality and the Lipschitz condition. In this section we demonstrate some applications of Theorem 1.5 to polyhedral multifunctions, which occur frequently in applications. A *polyhedral multifunction* is a multifunction whose graph is the union of a finite collection of convex polyhedral sets. If that collection consists of only a single set, then the multifunction is a *graph-convex polyhedral multifunction.* By [15, Proposition 1], if P is a polyhedral multifunction from $\mathbb{R}^n$ to $\mathbb{R}^m$, then there is some nonnegative number λ such that P is OLC with modulus λ at every point of $\mathbb{R}^n$. See also the result in [16, Example 9.57], which establishes, but does not state explicitly, the fact that a single modulus suffices, and which restricts the statement to $\operatorname{dom} P$.

The following corollaries combine Theorem 1.5 with the known OLC result for polyhedral multifunctions to derive results in forms convenient for application. We write $\operatorname{gph} F$ for the graph of a multifunction F.

COROLLARY 2.1. *Let F be a polyhedral multifunction from $\mathbb{R}^n$ to $\mathbb{R}^m$, let λ be its OLC modulus, and let C be a convex subset of* $\operatorname{dom} F$*. Then F is Lipschitzian with modulus λ on C if and only if it is inner semicontinuous relative to C at each point of C.*

Proof. As F is polyhedral, $\operatorname{gph} F$ is closed so F certainly has closed values, and we already know that F is everywhere OLC with modulus λ. Suppose first that F is inner semicontinuous relative to C at each point of C. By setting $X = C$ and applying Theorem 1.5, we conclude that F is Lipschitzian on C with modulus λ. On the other hand, if F is Lipschitzian on C with modulus λ, then Theorem 1.5 says that it must be inner semicontinuous relative to C at each point of C. ∎

The second corollary applies the first to a special case that is often easy to identify.

COROLLARY 2.2. *Let F be a polyhedral multifunction from $\mathbb{R}^n$ to $\mathbb{R}^m$, let λ be its OLC modulus, and let C be a convex subset of* $\operatorname{dom} F$ *on which F is single-valued. Then F is Lipschitzian on C with modulus λ.*

Proof. We need only show that F is inner semicontinuous relative to C at each point of C. Let x_0 be such a point, and let Q be any open set meeting $F(x_0)$. Because $F(x_0)$ is a singleton, we actually have $F(x_0) \in Q$, and because polyhedrality entails OLC there is a neighborhood V of x_0 such that if $x \in C \cap V$, then $F(x)$ is a singleton contained in Q. Therefore F is inner semicontinuous relative to C at x_0, and the claim then follows from Corollary 2.1. ∎

Our applications will involve variational inequalities posed over graph-convex polyhedral multifunctions. For a convex subset S of $\mathbb{R}^n$ and a point $x \in \mathbb{R}^n$, we use the symbol $N_S(x)$ to denote the *normal cone* of S at x, defined by

$$N_S(x) = \begin{cases} \{x^* \in \mathbb{R}^n \mid \text{for each } x' \in S,\ \langle x^*, x' - x\rangle \le 0\} & \text{if } x \in S, \\ \emptyset & \text{if } x \notin S. \end{cases}$$

The polyhedrality of the normal-cone operator N_S is well known (see, e.g., [7, p. 108], and the argument can be adapted with little difficulty to the case in which for a graph-convex polyhedral multifunction $S(u)$ we consider the multifunction sending (u, x) to $N_{S(u)}(x)$. This results in the following lemma.

LEMMA 2.3. *Let S be a graph-convex polyhedral multifunction from $\mathbb{R}^k$ to $\mathbb{R}^n$. Then the multifunction $G : \mathbb{R}^k \times \mathbb{R}^n \to \mathbb{R}^n$ defined by*

$$G(u, x) = N_{S(u)}(x) \tag{2.1}$$

is polyhedral.

The next proposition combines Lemma 2.3 with Corollaries 2.1 and 2.2 to produce a result about the Lipschitz continuity of solutions of an affine generalized equation posed over a polyhedral convex set, as functions of both the constant term in the generalized equation and the right-hand side of the constraints defining the polyhedral convex set.

PROPOSITION 2.4. *Let A be an $n \times n$ matrix and let S be a graph-convex polyhedral multifunction from $\mathbb{R}^k$ to $\mathbb{R}^n$. Define a multifunction $X : \operatorname{dom} S \times \mathbb{R}^n \to \mathbb{R}^n$ by*

$$X(u, v) = \{x \in \mathbb{R}^n \mid 0 \in Ax + v + N_{S(u)}(x)\}. \tag{2.2}$$

If C is a convex subset of $\operatorname{dom} X$, then X is Lipschitzian on C if and only if it is inner semicontinuous relative to C at each point of C. In particular, if X is single-valued on C, then it is necessarily Lipschitzian on C.

Proof. If we can establish that X is polyhedral, then the two claims will follow from Corollaries 2.1 and 2.2, respectively. Lemma 2.3 shows that the multifunction taking (u, x) to $N_{S(u)}(x)$ is polyhedral. As the sum of two polyhedral multifunctions is also polyhedral, the multifunction H taking (u, x) to $Ax + N_{S(u)}(x)$ is also polyhedral. However, a point (u, v, x) of $(\operatorname{dom} S) \times \mathbb{R}^n \times \mathbb{R}^n$ belongs to $\operatorname{gph} X$ if and only if $(u, x, -v) \in \operatorname{gph} H$. Therefore X is polyhedral. □

Special cases of Proposition 2.4 are well known, such as the Lipschitz continuity of the solution of a positive definite quadratic programming problem as a function of the vector in the quadratic objective function and the right-hand sides of the constraints defining the feasible set [18, Theorem 2.1], [3, Exercise 7.6.10].

3. Solutions of monotone affine variational inequalities. In 1975, Adler and Gale [1] gave an explicit representation of the set of solutions of a linear complementarity problem in $\mathbb{R}^n$ having a positive semidefinite matrix. Specifically, they exhibited a polyhedral convex set, depending only on the data of the complementarity problem, one of whose faces was the set of solutions of that problem. Similar representations appear in the work of Klatte and Thiere [8, p. 109] on convex quadratic programming and in the work of Luo and Tseng [12, Lemma 2] on affine variational inequalities with cocoercive operators. In section 3.1 we first extend the Adler–Gale results to the case of a monotone affine variational inequality posed over a polyhedral convex set. When that set is the nonnegative orthant, we recover the Adler–Gale

results. We then use an example in $\mathbb{R}^2$ to show that when we extend the situation to the more general case in which the set is a graph-convex polyhedral multifunction $S(u)$ depending on a parameter u, without further assumptions the solution set may behave very badly as u varies. If we want better behavior, we must impose stronger assumptions.

In section 3.2 we describe the well-known cocoercivity property of a monotone operator f; that is, the strong monotonicity of f^{-1}. As we are interested here in cases for which the monotone operator is affine, we discuss properties characterizing the class of matrices for which the cocoercivity property holds. These results come from the work of several investigators, and we briefly review the literature in this area. The class of cocoercive matrices includes, but is not limited to, all positive semidefinite symmetric matrices (thus all matrices arising in problems of linear, or of convex quadratic, programming).

We then point out that in the presence of the cocoercivity property two additional features appear, each of which will be important for our purposes. First, even for a nonlinear problem, the same normal vector appears in every solution of our problem. In our particular (polyhedral) application, we can show that this normal vector is Lipschitzian in the parameter u. Second, the solvability of the parametric variational inequality with polyhedral set $S(u)$ is independent of the value of u provided that it remains in $\operatorname{dom} S$. That is, if we write $X(u)$ for the multifunction expressing the dependence of the solution set on u, then $\operatorname{dom} X = \operatorname{dom} S$.

3.1. Positive semidefinite matrices. We will first extend the results of Adler and Gale [1] to the case of an affine variational inequality posed over a polyhedral convex set. Later we will want to make the set depend on a parameter, but the presence of the parameter is unnecessary here so we suppress it.

Let F and A be linear transformations from $\mathbb{R}^n$ to $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively, with A positive semidefinite. Let $f \in \mathbb{R}^m$ and $a \in \mathbb{R}^n$, and define S to be the polyhedral convex subset of $\mathbb{R}^n$ defined by $S = \{x \mid Fx \leq f\}$. Define X to be the (possibly empty) subset of $\mathbb{R}^n$ consisting of all points x, called *solutions,* satisfying

$$0 \in Ax + a + N_S(x). \tag{3.1}$$

For each solution x there is a unique point $x^* \in N_S(x)$ with $Ax + a + x^* = 0$. Because of the structure of N_S, we can represent x^* (generally not uniquely) in the form $x^* = F^*y^*$ with $y^* \in \mathbb{R}^m_+$ and $\langle y^*, Fx - f\rangle = 0$. Such y^* we call *multipliers associated with* x. For a solution x we write the set of all such multipliers y^* as $Y^*(x)$.

Now define a polyhedral convex subset Q of $\mathbb{R}^{n+m}$ by

$$Q = \{(x, y) \mid Fx \leq f,\ y^* \geq 0,\ Ax + a + F^*y^* = 0\}; \tag{3.2}$$

thus the definition of Q includes all conditions imposed on solutions and multipliers except the complementarity condition $\langle y^*, Fx - f\rangle = 0$. Define two subsets I and J of $\{1, \ldots, m\}$ by

$$I = \{i \in \{1, \ldots, m\} \mid \text{for each solution } x,\ (f - Fx)_i = 0\} \tag{3.3}$$

and

$$J = \{j \in \{1, \ldots, m\} \mid \text{for each multiplier } y^*,\ (y^*)_j = 0\}. \tag{3.4}$$

Finally, define Q_{IJ} by

$$Q_{IJ} = \{(x, y^*) \in Q \mid (f - Fx)_i = 0\ (i \in I),\ (y^*)_j = 0\ (j \in J)\}. \tag{3.5}$$

This Q_{IJ} is evidently a face of Q. The next theorem says that it consists precisely of solutions and their associated multipliers.

THEOREM 3.1. *One has*

$$Q_{IJ} = \{(x, y^*) \mid x \in X,\ y^* \in Y^*(x)\}. \tag{3.6}$$

Further, if x_0 and x_1 are elements of X with associated normal vectors $x_i^* = -(Ax_i + a)$ for $i = 0, 1$, then

$$\langle x_0^* - x_1^*, x_0 - x_1 \rangle = 0. \tag{3.7}$$

Proof. The proof has three main parts. We first show that the positive semidefiniteness of A implies (3.7) as well as a certain exchangeability of solution-multiplier pairs. Next, we use that information to show that each index in $\{1, \ldots, m\}$ is in I or in J (or possibly both), which we then use to prove (3.6).

Suppose that x_0 and x_1 are solutions, and let $y_i^* \in Y^*(x_i)$ for $i = 0, 1$. Then $x_i^* = F^* y_i^*$, $i = 0, 1$. The monotonicity of N_S and the positive semidefiniteness of A imply

$$0 \le \langle x_0^* - x_1^*, x_0 - x_1 \rangle = \langle -(Ax_0 - Ax_1), x_0 - x_1 \rangle \le 0, \tag{3.8}$$

which implies (3.7). Rewriting (3.7) as

$$0 = -\langle x_0^*, x_1 - x_0 \rangle - \langle x_1^*, x_0 - x_1 \rangle$$

and noting that each inner product is nonpositive because $x_i^* \in N_S(x_i)$ for each i, we see that each term is zero. Using $x_i^* = F^* y_i^*$ for $i = 0, 1$, we conclude that

$$0 = \langle x_0^*, x_1 - x_0 \rangle = \langle F^* y_0^*, x_1 - x_0 \rangle = \langle y_0^*, (Fx_1 - f) - (Fx_0 - f) \rangle = \langle y_0^*, Fx_1 - f \rangle,$$

as $\langle y_0^*, Fx_0 - f \rangle$ is zero by choice of y_0^*. Exchanging the roles of x_0 and x_1 yields $\langle y_1^*, Fx_0 - f \rangle = 0$.

Now choose any index $k \in \{1, \ldots, m\}$. If $k \notin I$, then there is some solution x with $(f - Fx)_k > 0$. If y^* is any multiplier associated with some solution x', then by what we have just shown we have $\langle y^*, Fx - f \rangle = 0$. As $y^* \ge 0$ and $Fx - f \le 0$, this implies that $(y^*)_k = 0$, so $k \in J$. Hence $I \cup J = \{1, \ldots, m\}$.

To prove (3.6), first suppose that $x \in X$ and $y^* \in Y^*(x)$. The definitions of I and J show that for each $i \in I$, $(Fx - f)_i = 0$ and for each $j \in J$, $(y^*)_j = 0$, so $(x, y) \in Q_{IJ}$.

Now choose $(x, y^*) \in Q_{IJ}$. The definition of Q shows that $Fx \le f$, so $x \in S$; also $y^* \ge 0$, and if we define $x^* = F^* y^*$, then $Ax + a + x^* = 0$. For each $i \in I$ we have $(Fx - f)_i = 0$ and for each $j \in J$ we have $(y^*)_j = 0$. As we have shown that $I \cup J = \{1, \ldots, m\}$, the inner product $\langle y^*, Fx - f \rangle$ must be zero, so that $x^* \in N_S(x)$. Therefore $x \in X$ and $y^* \in Y^*(x)$, which completes the proof of (3.6). ∎

In the case of a linear complementarity problem we have $F = -I$ and $f = 0$. Then (3.2) shows that we always have $y^* = Ax + a$, so that we can remove y^* from the problem and thus recover the Adler–Gale results.

Theorem 3.1 shows that the solution set of any monotone affine variational inequality over a polyhedral convex set is itself a polyhedral convex set, because it is the projection of a face of Q. We might wonder whether the information in the theorem could somehow be used to determine how this polyhedral convex solution set behaves

when one makes small changes in certain data of (3.1). For that purpose we now replace the set S by a graph-convex polyhedral multifunction defined for $u \in \mathbb{R}^k$ by

$$S(u) = \{x \mid Fx + Gu \leq g\}, \tag{3.9}$$

where G is a fixed $m \times k$ matrix and g is a fixed element of $\mathbb{R}^m$, and we consider the problem of finding x such that

$$0 \in Ax + a + N_{S(u)}(x). \tag{3.10}$$

For $u \in \mathbb{R}^k$ let $X(a, u)$ be the set of solutions of (3.1). This defines a multifunction $X : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^n$, whose values are solution sets of (3.10) for particular pairs (a, u).

From the form of (3.10) together with Lemma 2.3, we can see that X is a polyhedral multifunction, and hence by [15, Proposition 1] it is everywhere OLC with the same modulus. This guarantees stability with respect to expansion, as the set cannot expand at a rate faster than linear, but on the other hand there are no limits on its ability to contract. Indeed, one has only to think of a linear programming problem such as the trivial problem of maximizing ax for $x \in [0, 1] \subset \mathbb{R}$ with $a = 0$ initially, to see that the optimal set may contract disastrously for even slight changes in a. Luo and Tseng [10, Corollary 1] have shown how, by further restricting the class of matrices A, one may establish Lipschitz continuity with respect to a. However, in the rest of this paper we leave a fixed and write $X(u)$ for $X(a, u)$. This new X is then a multifunction taking values of u into solutions of (3.10), and we will seek conditions on the matrix A, beyond positive semidefiniteness, under which $X(\cdot)$ will actually be Lipschitzian (in the Pompeiu–Hausdorff metric) as opposed to outer Lipschitzian.

Klatte and Thiere [8, Theorem 4.2] have shown that the solution set of any convex quadratic programming problem is Lipschitzian as a function of the right-hand side vector of the constraints. They also refer to an earlier proof of that fact in [6]. Thus, the class of matrices A that we want to find must include those that are symmetric and positive semidefinite. We might hope that the class could include all positive semidefinite matrices, but to dismiss that possibility it is enough to consider the linear complementarity problem in $\mathbb{R}^2$ with constant vector zero and matrix

$$A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}; \tag{3.11}$$

that is,

$$0 \in Ax + N_{\mathbb{R}^2_+}(x),$$

where $\mathbb{R}^2_+$ is the nonnegative orthant of $\mathbb{R}^2$. We now parametrize this problem by defining, for $u \in \mathbb{R}$,

$$S(u) = \left\{ x \in \mathbb{R}^2 \mid (-I)x + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u \leq \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right\},$$

and consider the problem of solving $0 \in Ax + N_{S(u)}(x)$. If $u > 0$, then there is no solution, so $X(u) = \emptyset$; if $u = 0$, then $X(u) = \mathbb{R}_+ \times \{0\}$; if $u < 0$, then $X(u) = \{0\} \times [u, 0]$. Accordingly, for this problem $X(u)$ is not Lipschitzian in the Pompeiu–Hausdorff metric, nor even inner semicontinuous, although A is certainly positive semidefinite. Therefore, the class of matrices we seek must lie between the symmetric positive semidefinite matrices and the positive semidefinite matrices. We define this class in the next section and then develop some of its properties.

3.2. Cocoercivity. This section develops some properties of *cocoercive* operators that we need in the rest of this paper. We first define these operators, then give a theorem that characterizes cocoercivity for the specific class of operators that we use here.

DEFINITION 3.2. *Let T be a monotone operator from a Hilbert space H into H. We say that T is* cocoercive *if for some $\mu > 0$, T^{-1} is strongly monotone with modulus μ; that is, if for each (h,t) and (h',t') in the graph of T one has $\langle t'-t, h'-h\rangle \geq \mu\|t'-t\|^2$.*

Operators with strongly monotone inverses have received attention in the literature. In particular, Brézis and Haraux present a strong result about such operators that we use below [2, Observation (b), p. 175]. Also, it was observed in [14, Proposition 1] that even just strict monotonicity of T^{-1} implies a simple but useful fact, which we shall also use below. Suppose that M is another monotone operator from H into H, and consider the problem of finding solutions x of

$$0 \in T(x) + M(x). \tag{3.12}$$

If x_1 and x_2 are two such solutions, then there exist t_1 and t_2 such that for $i = 1, 2$ the pairs (x_i, t_i) and $(x_i, -t_i)$ belong to the graphs of T and M, respectively. Then

$$0 = \langle x_2 - x_1, t_2 - t_1\rangle + \langle x_2 - x_1, (-t_2) - (-t_1)\rangle.$$

Each inner product is nonnegative because T and M are monotone, so it follows that each is zero, and then the strict monotonicity of T^{-1} implies that $t_2 = t_1$. Accordingly, in this situation, if there are any solutions of (3.12) at all, then there is a *unique* t such that the solution set of (3.12) is $T^{-1}(t) \cap M^{-1}(-t)$. Luo and Tseng [12, Lemma 1] made a similar observation for affine functions on $\mathbb{R}^n$.

More recently, several authors have studied different aspects of such operators. Tseng [17] gave them the name *cocoercive,* which we use here. Other names include "strongly f-monotone" [13], "positive semidefinite plus" (or "psd-plus") for the case in which T is affine, referring to the property that $\langle x, Tx\rangle \geq 0$ implies $Tx = 0$, and the "Dunn property." Luo and Tseng [11] showed that a psd-plus linear operator $A : \mathbb{R}^n \to \mathbb{R}^n$ has a representation $A = E^T P E$, where P is a positive definite matrix of dimension r, where r is the rank of A. They also showed that for such an operator, if $A = C + K$ with C symmetric and K skew, then $\ker C \subset \ker K$. Interestingly, Iusem [5, Corollary 2] made a very similar observation for the quite different case of a symmetric copositive-plus matrix. Zhu and Marcotte [19, Proposition 2.5] showed that a linear operator A is psd-plus if and only if it is cocoercive. The book of Facchinei and Pang defines cocoercive operators [4, Volume 1, p. 79] and, in section 2.3 of Volume 1, discusses their properties including several mentioned above.

The following theorem records some useful facts about cocoercive linear operators on $\mathbb{R}^n$ in the form in which we need them in this paper. In it, we write $\ker A$ and $\operatorname{im} A$ for the kernel and image of a linear operator A, and we write A^{-1} for the inverse of A in the sense of multifunctions, not of linear algebra.

THEOREM 3.3. *Let A be a positive semidefinite linear operator from $\mathbb{R}^n$ to $\mathbb{R}^n$. Write $C = (A + A^*)/2$ and $K = (A - A^*)/2$ for the symmetric and skew parts of A, respectively. The following are equivalent:*

(a) *A is cocoercive with some positive modulus μ;*
(b) *A^{-1} is strictly monotone;*
(c) $\ker C \subset \ker K$.

If these equivalent properties hold, then one has $\ker A = \ker C = \ker A^*$ *and* $\operatorname{im} A = \operatorname{im} C = \operatorname{im} A^*$. *Finally,* A *is cocoercive if and only if* A^* *is cocoercive.*

Proof. (a) implies (b). This is obvious.

(b) implies (c). Suppose A^{-1} is strictly monotone. If $Cx = 0$, then as $\langle x, Kx\rangle = 0$ we have $0 = \langle x - 0, Ax - 0\rangle$. Strict monotonicity of A^{-1} applied to the pairs $(0, 0)$ and (Ax, x) yields $0 = Ax = Cx + Kx$, but as $Cx = 0$ we have $x \in \ker K$.

(c) implies (a). Suppose that $\ker C \subset \ker K$. We first observe that if $x \in \ker A$, then we have

$$0 = \langle x, Ax\rangle = \langle x, Cx\rangle + \langle x, Kx\rangle = \langle x, Cx\rangle,$$

and as C is positive semidefinite this implies $x \in \ker C$. Conversely, if $x \in \ker C$, then by hypothesis $x \in \ker K$ also, and then $x \in \ker A$. Thus if $\ker C \subset \ker K$, then we actually have $\ker A = \ker C$. But we can restate hypothesis (c) as $\ker C \subset \ker(-K)$, and then the argument that we just made shows that $\ker A^* = \ker C$. Therefore under this hypothesis we have $\ker A = \ker C = \ker A^*$, and by taking orthogonal complements we obtain also $\operatorname{im} A^* = \operatorname{im} C = \operatorname{im} A$.

If A is the zero operator, then it is cocoercive with any modulus, so assume $A \neq 0$; then the subspace $\operatorname{im} A^*$ has dimension at least 1. Define a function α from the nonempty compact set $W = \{x \in \operatorname{im} A^* \mid \|x\| = 1\}$ to $\mathbb{R}$ by $\alpha(x) = \langle x, Ax\rangle/\|Ax\|^2$. This α is well defined because $\operatorname{im} A^* = (\ker A)^\perp$, so that W contains no point with $Ax = 0$, and it is continuous on W; hence it takes a minimum there, say μ. If $\mu = 0$, then there is $x \in W$ with

$$0 = \langle x, Ax\rangle = \langle x, Cx\rangle + \langle x, Kx\rangle = \langle x, Cx\rangle,$$

so $x \in \ker C$. We have shown that then $x \in \ker A$ also, which contradicts the fact that $x \in (\operatorname{im} A^*)\backslash\{0\}$. Therefore $\mu > 0$.

Now choose any $x \in \mathbb{R}^n$. We can write x uniquely as $x = c + d$, with $c \in \ker A$ and $d \in \operatorname{im} A^*$. Then

$$\langle x, Ax\rangle = \langle c, Ac\rangle + \langle c, Ad\rangle + \langle d, Ac\rangle + \langle d, Ad\rangle = \langle d, Ad\rangle \geq \mu\|Ad\|^2 = \mu\|Ax\|^2,$$

where we used the fact that $\ker A = \ker A^*$. This proves the equivalence of (a), (b), and (c). We have already shown that (c) implies $\ker A = \ker C = \ker A^*$ and $\operatorname{im} A = \operatorname{im} C = \operatorname{im} A^*$. For the final assertion note that A is cocoercive if and only if $\ker C \subset \ker K$, which is the same as saying $\ker C \subset \ker(-K)$, which is equivalent to the cocoercivity of A^*. ∎

It follows from the last assertion of the theorem that when A is cocoercive one also has $\ker A = (\operatorname{im} A)^\perp$. Thus, the class of cocoercive linear operators shares some of the properties of the subclass consisting of symmetric positive semidefinite linear operators, though not all; the example

$$A = \begin{bmatrix} 1 & -1 \\ 1 & 2 \end{bmatrix} \tag{3.13}$$

shows that they need not be normal.

Such operators are also ubiquitous in applications, as they include, in particular, the linear transformations A appearing when (3.1) expresses a problem of linear programming (in which $A = 0$) or of convex quadratic programming (in which A is symmetric and positive semidefinite). However, (3.13) shows that these special cases do not exhaust the class.

Returning to the parametric problem (3.10), we assume that A is cocoercive with some modulus $\mu > 0$. Then the uniqueness result from [14, Proposition 1] mentioned previously tells us that there is a function $x^* : \operatorname{dom} S \to \mathbb{R}^n$ such that for each $u \in \operatorname{dom} S$,

$$X(u) = N_{S(u)}^{-1}[x^*(u)] \cap \{x \mid Ax + a + x^*(u) = 0\}. \tag{3.14}$$

The set $N_{S(u)}^{-1}[x^*(u)]$ is the maximal face F of $S(u)$ on which $x^*(u)$ is everywhere a normal vector. As $\{x \mid Ax + a + x^*(u) = 0\}$ is a face of itself, and as the faces of an intersection of two convex sets are the intersections of their faces, $X(u)$ is a face of the polyhedral convex set

$$P(u) = S(u) \cap \{x \mid Ax + a + x^*(u) = 0\}. \tag{3.15}$$

Luo and Tseng [12, Lemma 2] gave a similar facial characterization of the optimal set of a cocoercive affine variational inequality, but for the case of a fixed underlying set and with a different proof.

The sets $X(u)$ and $P(u)$ play roles similar to those of $Q(u)$ and $Q_{IJ}(u)$ defined by (3.2) and (3.5), respectively, but with S replaced by the $S(u)$ defined in (3.9). However, they are simpler than $Q(u)$ and $Q_{IJ}(u)$, because the availability of $x^*(u)$ has enabled us to decouple the solutions x from the multipliers y^*; indeed, no multipliers appear in (3.15). This is entirely due to the cocoercivity assumption, and it is the crucial step in the analysis because without making very restrictive assumptions on F (such as a linear independence condition), we cannot control the behavior of the multipliers.

However, in order to make this new formulation tractable we have to establish good behavior of $x^*(u)$. We do this in two steps: First we show that under cocoercivity and a solvability condition $\operatorname{dom} X = \operatorname{dom} S$, and then we show that $x^*(\cdot)$ is Lipschitzian on that set.

We first observe that for each $u \in \operatorname{dom} S$, $\operatorname{im} N_{S(u)} = F^*(\mathbb{R}^m_+)$. Indeed, the form of $S(u)$ shows that $\operatorname{im} N_{S(u)} \subset F^*(\mathbb{R}^m_+)$. On the other hand, for any $y^* \in \mathbb{R}^m_+$ the objective value in the linear programming problem

$$\sup\{\langle F^* y^*, x\rangle \mid x \in S(u)\}$$

is bounded above by $\langle y^*, g - Gu\rangle$, and therefore the problem actually has a solution x'. Then $F^* y^* \in N_{S(u)}(x')$, so $F^*(\mathbb{R}^m_+) \subset \operatorname{im} N_S(u)$.

PROPOSITION 3.4. *Let $T(u,x) = Ax + a + N_{S(u)}(x)$. If A is cocoercive, then for each $u \in \operatorname{dom} S$ one has*

$$\operatorname{im} T(u,\cdot) = \operatorname{im} A + a + F^*(\mathbb{R}^m_+). \tag{3.16}$$

Proof. The operator $T(u,\cdot)$ is the sum of the two maximal monotone operators $A(\cdot)+a$ and $N_{S(u)}$. As the relative interiors of the domains of these two operators meet, the sum is also maximal monotone. As shown in [2, Proposition 2], the first operator satisfies the key condition $(*)$ of that work because it is cocoercive. Moreover, its domain includes that of the second operator. By [2, Theorem 4], the image of $T(u,\cdot)$ is then "almost equal" to the sum of the images of the two constituent operators, in the sense that the closures and the interiors of these two sets are equal. The operators being polyhedral, their images are unions of finite collections of polyhedral convex sets; hence each is closed and so is their sum. Therefore (3.16) holds. □

The expression on the right in (3.16) is independent of u. Hence we have the following corollary.

COROLLARY 3.5. *If A is cocoercive and*

$$0 \in \operatorname{im} A + a + F^*(\mathbb{R}^m_+), \tag{3.17}$$

then $\operatorname{dom} X = \operatorname{dom} S$.

Proof. Evidently $\operatorname{dom} X \subset \operatorname{dom} S$. If $u \in \operatorname{dom} S$, then the hypotheses together with (3.16) show that $0 \in \operatorname{im} T(u, \cdot)$, so that $u \in \operatorname{dom} X$ and so $\operatorname{dom} S \subset \operatorname{dom} X$. ∎

If the condition in (3.17) does not hold, then (3.10) has no solution for any u, whereas if it holds, then (3.10) is solvable for every $u \in \operatorname{dom} S$. In what follows we refer to (3.17) as the solvability condition.

Finally, we show that the unique normal vector $x^*(u)$ defined by (3.14) is Lipschitzian in u.

PROPOSITION 3.6. *If A is cocoercive and $0 \in \operatorname{im} A + a + F^*(\mathbb{R}^m_+)$, then $x^*(\cdot)$ is Lipschitzian on* $\operatorname{dom} S$.

Proof. The graph of $x^*(\cdot)$ is

$$\begin{aligned}\operatorname{gph} x^*(\cdot) = \{(u, x^*) \in \mathbb{R}^k \times \mathbb{R}^n \mid &\text{for some } x \in \mathbb{R}^n,\\ &Ax + a + x^* = 0 \text{ and } (x, x^*) \in N_{S(u)}\}.\end{aligned}$$

This set is the projection into $\mathbb{R}^k \times \mathbb{R}^n$ of the set

$$\{(u, x, x^*) \in \mathbb{R}^k \times \mathbb{R}^n \times \mathbb{R}^n \mid Ax + a + x^* = 0,\ (x, x^*) \in N_{S(u)}\},$$

which is a union of polyhedral convex sets; hence so is the graph of $x^*(\cdot)$, which is therefore a polyhedral multifunction. The domain of $x^*(\cdot)$ is $\operatorname{dom} X$, which under our hypotheses is $\operatorname{dom} S$. But we showed just after (3.12) that for each u there could be no more than one x^*, and if $u \in \operatorname{dom} X$, then there must be at least one x^*. Therefore $x^*(\cdot)$ is single-valued on the convex set $\operatorname{dom} S$. By Corollary 2.2, it is then Lipschitzian on $\operatorname{dom} S$. ∎

4. Lipschitz continuity of solution sets. In this section we apply the work of the preceding sections to show that if the assumptions of cocoercivity and solvability hold, then the solution multifunction $X(\cdot)$ of the problem (3.10) is Lipschitzian on $\operatorname{dom} S$ in the Pompeiu–Hausdorff metric. We then show that for any positive semidefinite A that is not cocoercive, there exists a graph-convex polyhedral multifunction S such that the problem (3.10) with $a = 0$ is solvable, but has a solution set that not only fails to be Lipschitzian, but fails even to be inner semicontinuous. Therefore cocoercivity is the weakest assumption that we can impose on A to ensure Lipschitz continuity of $X(\cdot)$.

THEOREM 4.1. *If A is cocoercive and $0 \in \operatorname{im} A + a + F^*(\mathbb{R}^m_+)$, then X is Lipschitzian in the Pompeiu–Hausdorff metric on* $\operatorname{dom} S$.

Proof. As we observed just after (3.10), X is a polyhedral multifunction, and therefore there is some modulus λ such that X is everywhere OLC with modulus λ. We will show that under our hypotheses, X is inner semicontinuous relative to $\operatorname{dom} S$ at each $u_0 \in \operatorname{dom} S$, and Theorem 1.5 will then establish that it is Lipschitzian as claimed.

Choose some u_0 in $\operatorname{dom} S$, which by Corollary 3.5 is also $\operatorname{dom} X$. Let Q be an open set meeting $X(u_0)$, and choose some $x_0 \in Q \cap X(u_0)$. The technique of the proof is to construct a finite number of graph-convex polyhedral multifunctions such that for each u near u_0, one of these will yield a point in the set $X(u)$ that is close to x_0.

Let L be the set of indices i in $\{1,\dots,m\}$ for which the inequality $(Fx_0+Gu_0)_i \le g_i$ is satisfied as an equality. Write cL for the complement of L in $\{1,\dots,m\}$. We can find neighborhoods U_0 of u_0 and V_0 of x_0, with $V_0 \subset Q$, so that whenever $(u,x) \in U_0 \times V_0$ the inequality $(Fx+Gu)_i \le g_i$ is strict for each $i \in cL$. Find a positive ϵ so that the ball $x_0 + \epsilon B$ lies in V_0.

Now for $u \in \operatorname{dom} S$ denote by $X_L(u)$ the solution set of the reduced problem

$$0 \in Ax + a + N_{S_L(x)}, \tag{4.1}$$

where $S_L(u) = \{x \in \mathbb{R}^n \mid F_L x + G_L u \le g_L\}$, and where we write F_L for the submatrix of F whose rows belong to the index set L, and similarly for G_L and g_L. If $L = \emptyset$, then we take $S_L(u) = \mathbb{R}^n$. We know the problem in (4.1) is solvable for $u = u_0$, because x_0 solves it. Therefore this reduced problem satisfies the solvability condition, as well as the cocoercivity condition. Accordingly, Corollary 3.5 shows that $\operatorname{dom} X_L = \operatorname{dom} S_L \supset \operatorname{dom} S$, where the inclusion holds because the reduced problem has fewer constraints than did the original one. Also, Proposition 3.6 shows that the function $x_L^*(\cdot)$ appearing in the reduced problem is Lipschitzian with some modulus ξ_L.

Now for all possible partitions (I,J) of L, consider the graph-convex polyhedral multifunction

$$P_{I,J}(b,c,d) = \{x \in \mathbb{R}^n \mid F_I x = b,\ F_J x \le c,\ Ax = d\}, \tag{4.2}$$

where $(b,c,d) \in \mathbb{R}^{|I|} \times \mathbb{R}^{|J|} \times \mathbb{R}^n$. Each of these multifunctions is Lipschitzian on its domain, and there are finitely many of them; let λ be the maximum of their Lipschitz constants. Write $s_L(u)$ for $g_L - G_L u$, and find a neighborhood U_1 of u_0, contained in U_0, such that whenever $u \in U_1 \cap \operatorname{dom} S$ we have

$$\lambda \left\| \begin{bmatrix} s_L(u) \\ -x_L^*(u) \end{bmatrix} - \begin{bmatrix} s_L(u_0) \\ -x_L^*(u_0) \end{bmatrix} \right\| < \epsilon. \tag{4.3}$$

Now choose any $u \in U_1 \cap \operatorname{dom} S$. Applying the discussion at (3.15) to the reduced problem, we see that $X_L(u)$ is a face of $P_L(u)$. Accordingly, there will be some partition I, J of L for which

$$X_L(u) = P_{I,J}(s_I(u), s_J(u), -[x_L^*(u) + a]). \tag{4.4}$$

We also have

$$x_0 \in P_{I,J}(s_I(u_0), s_J(u_0), -[x_L^*(u_0) + a]). \tag{4.5}$$

If we write $M(u)$ for $P_{I,J}(s_I(u), s_J(u), -[x_L^*(u)+a])$, then the Lipschitz continuity of $P_{I,J}$ in the Pompeiu–Hausdorff metric yields

$$M(u_0) \subset M(u) + \lambda \left\| \begin{bmatrix} s_L(u) \\ -x_L^*(u) \end{bmatrix} - \begin{bmatrix} s_L(u_0) \\ -x_L^*(u_0) \end{bmatrix} \right\| B \subset M(u) + \epsilon B.$$

As $M(u) = X_L(u)$ and $M(u_0)$ contains x_0, this implies that there is a point $x \in X_L(u)$ such that $x \in x_0 + \epsilon B$. In particular, $x \in Q$. Also, the pair (u,x) belongs to $U_0 \times V_0$, so every constraint $F_i x + G_i u \le g_i$ with index $i \in cL$ is slack there. That means that not only is $x_L^*(u)$ in $N_{S_L(u)}(x)$, but it also belongs to $N_{S(u)}(x)$, and therefore $x \in X(u)$. Therefore $X(u)$ is inner semicontinuous at u_0 relative to $\operatorname{dom} S$, which completes the proof. □

Theorem 4.1 shows that if cocoercivity and the solvability condition hold, then the solution set of (3.10) is Lipschitzian on its domain. The last theorem will show that if A is positive semidefinite but not cocoercive, then one can find a graph-convex polyhedral multifunction S so that the problem (3.10) is solvable, but its solution multifunction X is not inner semicontinuous, hence *a fortiori* not Lipschitzian. In fact, $S(u)$ can be taken to be a translated halfline.

THEOREM 4.2. *Let A be a positive semidefinite linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^n$ that is not cocoercive. Then there exists a graph-convex polyhedral multifunction $S : \mathbb{R} \to \mathbb{R}^n$ such that the problem of finding x such that $0 \in Ax + N_{S(u)}(x)$ is solvable for $u_0 = 0$, but there are points u arbitrarily close to u_0 such that $X(u) \neq \emptyset$ but the inclusion $X(u_0) \subset X(u) + \alpha B$ holds for no real number α.*

Proof. As A is not cocoercive, by Theorem 3.3 its inverse is not strictly monotone. Therefore there is some $x \in \mathbb{R}^n$ such that $\langle x, Ax\rangle = 0$ but $Ax \neq 0$. Write $A = C + K$ with C symmetric and K skew. As $\langle x, Ax\rangle = 0$ we have $\langle x, Cx\rangle = 0$ and therefore $Cx = 0$, which with $Ax \neq 0$ implies that $Kx \neq 0$. As $A^* = C - K$, it follows that the point $y = A^*x$ is not zero. For $u \in \mathbb{R}$ define

$$S(u) = uy + (\mathbb{R}_+)x.$$

The graph of S is

$$\{(u, uy + \xi x) \mid u \in \mathbb{R}, \xi \geq 0\} = \{u(1, y) + \xi(0, x) \mid u \in \mathbb{R}, \xi \geq 0\},$$

which is a polyhedral convex set. Now fix $u \in \mathbb{R}$ and $\xi \geq 0$, and let $x(u, \xi) = uy + \xi x$. Then we have

$$\langle x, -Ax(u, \xi)\rangle = -u\|y\|^2.$$

For $u < 0$ this quantity is positive, so $-Ax(u, \xi)$ is not in $N_{S(u)}(x(u, \xi))$ for any $\xi \geq 0$; thus $X(u) = \emptyset$. For $u = 0$, $-u\|y\|^2 = 0$, so $-Ax(u, \xi) \in N_{S(u)}(x(u, \xi))$ for each $\xi \geq 0$, and $X(u)$ is the entire halfline $S(u)$. For $u > 0$ we have $-u\|y\|^2 < 0$, so $-Ax(u, \xi) \in N_{S(u)}(x(u, \xi))$ for $\xi = 0$ but for no $\xi > 0$, so $X(u) = \{uy\}$. Therefore, for $u > 0$ the set $X(u)$ is a singleton, and therefore the halfline $X(0)$ contains points arbitrarily far away from $X(u)$. ∎

5. Appendix: Proof of Theorem 1.5. If X is empty there is nothing to prove, so we assume X to be nonempty.

(a) *implies* (b). Choose two points of X and call them x_0 and x_1. For each $t \in (0, 1)$ let $x_t = (1 - t)x_0 + tx_1$. The OLC hypothesis ensures that for each $t \in [0, 1]$ there is a ball $B(x_t, \rho_t)$ about x_t with positive radius ρ_t such that for each x' in $X \cap B(x_t, \rho_t)$, $S(x') \subset S(x_t) + \lambda\|x' - x_t\|B$. Define

$$\tau = \sup\{t \in [0, 1] \mid \text{for each } s \in [0, t],\ S(x_s) \subset S(x_0) + \lambda\|x_s - x_0\|B\}.$$

We have $\tau > 0$ because ρ_0 is positive. We show first that

$$S(x_\tau) \subset S(x_0) + \lambda\|x_\tau - x_0\|B. \tag{5.1}$$

By assumption the set $S(x_0)$ is closed, and therefore so is $S(x_0) + \lambda\|x_\tau - x_0\|B$; let Q be the complement of the latter set. If (5.1) were not true, then $S(x_\tau)$ would meet the open set Q, and the inner semicontinuity hypothesis would then imply the existence

of $\sigma \in [0, \tau)$ for which $S(x_\sigma)$ also met Q. This cannot be true, because all such σ satisfy

$$S(x_\sigma) \subset S(x_0) + \lambda \|x_\sigma - x_0\| B \subset S(x_0) + \lambda \|x_\tau - x_0\| B.$$

This establishes (5.1).

If τ were less than 1 there would be $\kappa \in (\tau, 1)$ with $\|x_\kappa - x_\tau\| < \rho_\tau$, such that

$$S(x_\kappa) \not\subset S(x_0) + \lambda \|x_\kappa - x_0\| B. \tag{5.2}$$

However, we would then have from (5.1) and the definition of ρ_τ

$$\begin{aligned} S(x_\kappa) &\subset S(x_\tau) + \lambda \|x_\kappa - x_\tau\| B \\ &\subset S(x_0) + \lambda (\|x_\kappa - x_\tau\| + \|x_\tau - x_0\|) B \\ &= S(x_0) + \lambda \|x_\kappa - x_0\|) B, \end{aligned}$$

where the equality in the last line holds because x_0, x_τ, and x_κ are collinear. This contradicts (5.2), so τ must be 1. Putting $\tau = 1$ in (5.1) shows that $S(x_1) \subset S(x_0) + \lambda \|x_1 - x_0\| B$, and reversing the roles of x_0 and x_1 completes the proof.

(b) *implies* (a). If (b) holds, then S is *a fortiori* OLC relative to X at each point $x \in X$ with modulus λ. For inner semicontinuity, let $x \in X$ and let Q be an open set meeting $S(x)$. Then there exist a point $y \in S(x) \cap Q$ and a ball $B(y, \eta)$ with $\eta > 0$ such that $B(y, \eta) \subset Q$. Let ν be a positive number with $\lambda \nu \leq \eta$, and take V to be $X \cap B(x, \nu)$. If $x' \in V$, then (b) yields

$$y \in S(x) \subset S(x') + \lambda \|x' - x\| B \subset S(x') + \eta B.$$

This means that there is some $y' \in S(x')$ with $\|y' - y\| \leq \eta$, so that $y' \in B(y, \eta)$, and hence $y' \in S(x') \cap Q$. Therefore $S(x')$ meets Q, so S is inner semicontinuous at x relative to X. □

Acknowledgments. I wish to thank Shu Lu, Paul Tseng, and two anonymous referees for comments and criticisms that have greatly improved this paper.

REFERENCES

[1] I. Adler and D. Gale, *On the Solutions of the Positive Semi-Definite Complementarity Problem*, Technical report ORC 75-12, Operations Research Center, University of California, Berkeley, Berkeley, CA, 1975.

[2] H. Brézis and A. Haraux, *Image d'une somme d'operateurs monotones et applications*, Israel J. Math., 23 (1976), pp. 165–186.

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

[4] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Springer Series in Operations Research, Springer-Verlag, New York, 2003. Published in two volumes, paginated continuously.

[5] A. N. Iusem, *On the convergence of iterative methods for symmetric linear complementarity problems*, Math. Programming, 59 (1993), pp. 33–48.

[6] D. Klatte, *Beiträge zur Stabilitätsanalyse nichtlinearer Optimierungsprobleme*. Dissertation B (Habilitationsschrift), Sektion Mathematik, Humboldt-Universität Berlin, Berlin, Germany, 1984.

[7] D. Klatte and B. Kummer, *Nonsmooth Equations in Optimization*, Regularity, Calculus, Methods and Applications, Kluwer Academic Publishers, Dordrecht, the Netherlands, 2002.

[8] D. Klatte and G. Thiere, *Error bounds for solutions of linear equations and inequalities*, ZOR—Math. Methods Oper. Res., 41 (1995), pp. 191–214.

[9] W. Li, *Sharp Lipschitz constants for basic optimal solutions and basic feasible solutions of linear programs*, SIAM J. Control Optim., 32 (1994), pp. 140–153.
[10] X.-D. Luo and P. Tseng, *On a global projection-type error bound for the linear complementarity problem*, Linear Algebra Appl., 253 (1997), pp. 251–278.
[11] Z.-Q. Luo and P. Tseng, *A decomposition property for a class of square matrices*, Appl. Math. Lett., 4 (1991), pp. 67–69.
[12] Z.-Q. Luo and P. Tseng, *On a global error bound for a class of monotone affine variational inequality problems*, Oper. Res. Lett., 11 (1992), pp. 159–165
[13] T. L. Magnanti and G. Perakis, *The orthogonality theorem and the strong-f-monotonicity condition for variational inequality algorithms*, SIAM J. Optim., 7 (1997), pp 248–273.
[14] S. M. Robinson, *Inverse sums of monotone operators*, Game Theory and Mathematical Economics, O. Moeschlin and D. Pallaschke, eds., North-Holland, Amsterdam, 1981, pp. 449–457.
[15] S. M. Robinson, *Some continuity properties of polyhedral multifunctions*, Math. Programming Stud., 14 (1981), pp. 206–214.
[16] R. T. Rockafellar and Roger J–B Wets, *Variational Analysis*, Grundlehren der Mathematischen Wissenschaften 317, Springer-Verlag, Berlin, 1998.
[17] P. Tseng, *Further applications of a splitting algorithm to decomposition in variational inequalities and convex programming*, Math. Programming, 48 (1990), pp. 249–263.
[18] N. D. Yen, *Lipschitz continuity of solutions of variational inequalities with a parametric polyhedral constraint*, Math. Oper. Res., 20 (1995), pp. 695–708.
[19] D. L. Zhu and P. Marcotte, *Co-coercivity and its role in the convergence of iterative schemes for solving variational inequalities*, SIAM J. Optim., 6 (1996), pp. 714–726.

SIAM J. OPTIM.
Vol. 18, No. 3, pp. 1061–1081

A NEW CLASS OF ALTERNATING PROXIMAL MINIMIZATION ALGORITHMS WITH COSTS-TO-MOVE*

H. ATTOUCH†, P. REDONT†, AND A. SOUBEYRAN‡

Abstract. Given two objective functions $f : \mathcal{X} \mapsto \mathbb{R} \cup \{+\infty\}$ and $g : \mathcal{Y} \mapsto \mathbb{R} \cup \{+\infty\}$ on abstract spaces $\mathcal{X}$ and $\mathcal{Y}$, and a coupling function $c : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}^+$, we introduce and study alternative minimization algorithms of the following type: $(x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$ given; $(x_n, y_n) \to (x_{n+1}, y_n) \to (x_{n+1}, y_{n+1})$ as follows:
$\begin{cases} x_{n+1} \in \operatorname{argmin}\{f(\xi) + \beta_n c(\xi, y_n) + \alpha_n h(x_n, \xi) : \ \xi \in \mathcal{X}\}, \\ y_{n+1} \in \operatorname{argmin}\{g(\eta) + \mu_n c(x_{n+1}, \eta) + \nu_n k(y_n, \eta) : \ \eta \in \mathcal{Y}\}. \end{cases}$
Their most original feature is the introduction of the terms $h : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}^+$ and $k : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}^+$ which are costs to change or to move (distance-like functions, relative entropies) accounting for various inertial, friction, or anchoring effects. These algorithms are studied in a general abstract framework. The introduction of the costs to change h and k leads to proximal minimizations with corresponding dissipative effects. As a result, the algorithms enjoy nice convergent properties. Coefficients α_n, β_n, μ_n, ν_n are nonnegative parameters. When taking $\alpha_n = \nu_n = 0$ and quadratic costs on a Hilbert space, one recovers the classical alternating minimization algorithm, which itself is a natural extension of the alternating projection algorithm of von Neumann. A number of new significant results hold in general metric spaces. We pay particular attention to the following cases: (1.) $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ are complete metric spaces and $h \geq d_{\mathcal{X}}$, $k \geq d_{\mathcal{Y}}$ ("high local costs to move"); the algorithms then provide sequences that converge to Nash equilibria. (2.) $\mathcal{X} = \mathcal{Y} = \mathcal{H}$ is a Hilbert space, the costs to change are quadratic ("low local costs to move") and the functions $f, g : \mathcal{H} \mapsto \mathbb{R} \cup \{+\infty\}$ are closed, convex, proper; then some of the classical convergence theorems for alternating convex minimization algorithms, including those of Acker and Prestel, are properly extended with original proofs.

Key words. alternating minimization, alternating projection, proximal algorithms, costs to change, inertia, anchoring effect, dynamical game theory, Nash equilibria, steepest descent, dissipative dynamical systems

AMS subject classifications. 65K05, 49J40, 49M45, 90B50, 90C25, 90C29, 90D50, 92J10

DOI. 10.1137/060657248

1. Introduction and general presentation.

1.1. Some classical results about alternating minimization. Let us first recall some basic facts about alternating minimization and proximal algorithms that will be most useful to throw light on the original aspects of our approach.

(a) The starting fundamental result is due to von Neumann (1950) [33]. Let H be a Hilbert space and let C_1, C_2 be two closed affine subspaces of H with $C_1 \cap C_2 \neq \emptyset$. Let P_{C_1} and P_{C_2} denote the orthogonal projections on C_1 and C_2. Then, for any x_0 in H, the sequence $(x_n)_{n \in \mathbb{N}}$ obtained by alternatively projecting on C_1 and on C_2, namely

$$x_n = (P_{C_1} \circ P_{C_2})^n x_0,$$

*Received by the editors April 14, 2006; accepted for publication (in revised form) June 6, 2007; published electronically October 4, 2007.
http://www.siam.org/journals/siopt/18-3/65724.html

†Institut de Mathématiques et de Modélisation de Montpellier, UMR CNRS 5149, CC 51, Université Montpellier II, Place Eugène Bataillon, 34095 Montpellier cedex 5, France (attouch@math.univ-montp2.fr, redont@math.univ-montp2.fr). These authors acknowledge the support of the French ANR under grant ANR-05-BLAN-0248-01.

‡GREQAM UMR CNRS 6579, Université de la Méditerranée, 13290 Les Milles, France (soubey@romarin.univ-aix.fr).

strongly converges to the projection of x_0 onto $C_1 \cap C_2$. This result, and its extension by Halperin [24] to the case of cyclic projections onto a finite number of closed affine subspaces, provides a powerful tool for solving convex feasibility problems in Hilbert spaces. A rich literature has been devoted to this subject; see, for example, the extensive studies by Deutsch [23], Bauschke, Borwein, and Lewis [14], and Combettes [22].

(b) A next decisive step in the understanding of the convergence analysis of these algorithms in the general framework of convex optimization has been made by Acker and Prestel [1]. Let $f, g : H \mapsto \mathbb{R} \cup \{+\infty\}$ be two closed convex proper functions on the Hilbert space H. Fix $(x_0, y_0) \in H \times H$ and consider the sequence generated by the alternating minimization algorithm

$$\begin{cases} x_{n+1} = \operatorname{argmin}\{f(x) + \frac{1}{2} \parallel x - y_n \parallel^2 : x \in H\}, \\ y_{n+1} = \operatorname{argmin}\{g(y) + \frac{1}{2} \parallel x_{n+1} - y \parallel^2 : y \in H\}. \end{cases}$$

Then, the sequence $(x_n, y_n)_{n\in\mathbb{N}}$ weakly converges to a solution of the joint minimization problem on $H \times H$,

$$\min \left\{ f(x) + g(y) + \frac{1}{2} \parallel x - y \parallel^2 : (x, y) \in H \times H \right\},$$

if we assume that the minimum point set is nonempty. This result can be formulated in an equivalent form by using the proximal mappings introduced by Moreau [32] and Rockafellar [36]: Recall that, for a closed convex proper function $\varphi : H \mapsto \mathbb{R} \cup \{+\infty\}$, the proximal mapping $\operatorname{prox}_\varphi$ is defined by

$$\forall z \in H, \quad \operatorname{prox}_\varphi(z) = \operatorname{argmin}_\xi \left\{ \varphi(\xi) + \frac{1}{2} \parallel z - \xi \parallel^2 \right\}.$$

When $\varphi = \delta_C$ is the indicator function of a closed convex nonempty set C, one has $\operatorname{prox}_\varphi = P_C$, the classical projection, whence the terminology. Like projections, proximal mappings are nonexpansive (more precisely they are firmly nonexpansive). Acker and Prestel's algorithm may be reformulated as

$$\begin{cases} x_{n+1} = \operatorname{prox}_f y_n, \\ y_{n+1} = \operatorname{prox}_g x_{n+1}. \end{cases}$$

In particular, it provides a nice extension of von Neumann's theorem to two closed convex nonempty sets of the Hilbert space H (take $f = \delta_{C_1}$, $g = \delta_{C_2}$); namely the following.

1. If $C_1 \cap C_2 \neq \emptyset$, then the sequences x_n, y_n generated by the alternating projection algorithm weakly converge,

$$w - \lim x_n = w - \lim y_n = \overline{z},$$

to a point $\overline{z} \in C_1 \cap C_2$.

2. If $C_1 \cap C_2 = \emptyset$, then $w - \lim x_n = \overline{x}$ and $w - \lim y_n = \overline{y}$ exist where $\overline{x} \in C_1$ and $\overline{y} \in C_2$ are such that $\parallel \overline{x} - \overline{y} \parallel$ achieves the distance between sets C_1 and C_2,

$$\parallel \overline{x} - \overline{y} \parallel = \inf\{\parallel x - y \parallel : x \in C_1,\, y \in C_2\}.$$

This last result has proved to be of fundamental importance for applications, especially when solving inverse problems (possibly ill-posed) arising in various fields of engineering (mechanics, signal reconstruction, image processing, statistics, etc.). Indeed, in

the case of inconsistent constraints ($C_1 \cap C_2 = \emptyset$) due to inaccurate measurements, for example, the previous result guarantees the convergence of the algorithm to a relaxed solution $(\overline{x}, \overline{y})$ taking the inconsistent constraints into account in the best possible way; see [22, 14].

Note that, and this will be important for further developments, the preceding algorithm can be viewed as the alternating minimization of the bivariate function

$$L : (x, y) \in H \times H \mapsto L(x, y) = f(x) + g(y) + \frac{1}{2} \parallel x - y \parallel^2 \in \mathbb{R} \cup \{+\infty\}$$

and provides a minimizing sequence for L. A related approach has recently been developed by Bauschke, Combettes, and Noll [15] who consider the alternating minimization procedure in the case of the Euclidean space for the bivariate function

$$L(x, y) = f(x) + g(y) + D(x, y),$$

where $D(x, y)$ is a coupling function of Bregman type, namely

$$D(x, y) = \theta(x) - \theta(y) - \langle \nabla\theta(y), x - y \rangle,$$

with θ a convex function designed for the developments of interior point methods in convex programming (note that $D(x, y) = 1/2 \parallel x - y \parallel^2$ when $\theta(x) = 1/2 \parallel x \parallel^2$).

(c) To complete this general portrait of alternating proximal methods, let us mention the approach via the asymptotic analysis of the composition of resolvents of maximal monotone operators. This allows us to treat in a unified way the case of convex functions, min-max, and complementary problems. The corresponding generated semigroups of contractions and the Trotter–Kato–Lie formula make a natural link with dynamical systems [16, 35, 19, 30].

1.2. Description of the alternating algorithms. Due to the wide potential range of applications of these algorithms (from engineering to decision sciences) we adopt a quite general terminology (the corresponding terminology in game theory is described in section 1.4).

$\mathcal{X}$ and $\mathcal{Y}$ are abstract spaces.

f: $\mathcal{X} \mapsto \mathbb{R} \cup \{+\infty\}$ is the "first" criterion or objective function.

g: $\mathcal{Y} \mapsto \mathbb{R} \cup \{+\infty\}$ is the "second" criterion or objective function.

Nonnegative bivariate functions h, k, c account for the following effects:

h: $(x, \xi) \in \mathcal{X} \times \mathcal{X} \mapsto h(x, \xi) \in \mathbb{R}^+ \cup \{+\infty\}$ is the cost to change from $x \in \mathcal{X}$ to $\xi \in \mathcal{X}$, which is involved in the minimization procedure of f on $\mathcal{X}$. For example, $h(x, \xi) = d_{\mathcal{X}}(x, \xi)$ in the case of a metric space $(\mathcal{X}, d_{\mathcal{X}})$.

k: $(y, \eta) \in \mathcal{Y} \times \mathcal{Y} \mapsto k(y, \eta) \in \mathbb{R}^+ \cup \{+\infty\}$ similarly is the cost to change from $y \in \mathcal{Y}$ to $\eta \in \mathcal{Y}$, which is involved in the minimization procedure of g on $\mathcal{Y}$.

For example, functions h and k may account for physical costs in transportation processes.

The function $c : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}^+$ couples the two problems. For example, in dynamical games (see section 1.4) it may account for attracting or repulsive coupling effects between two players.

We can now describe the algorithm:

- $(x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$ given is the initial state;
- $(x_n, y_n) \mapsto (x_{n+1}, y_n) \mapsto (x_{n+1}, y_{n+1})$ as follows:

$$(\mathrm{CA}^2) \begin{cases} x_{n+1} \in \operatorname{argmin}\{f(\xi) + \beta_n c(\xi, y_n) + \alpha_n h(x_n, \xi) : \ \xi \in \mathcal{X}\}, \\ y_{n+1} \in \operatorname{argmin}\{g(\eta) + \mu_n c(x_{n+1}, \eta) + \nu_n k(y_n, \eta) : \eta \in \mathcal{Y}\}, \end{cases}$$

where α_n, β_n, μ_n, ν_n are nonnegative parameters. We call (CA^2) the cognitive alternating algorithm for two criteria or two agents.

Indeed, for numerical purposes as well as for the realism of our model, it is convenient to introduce an approximate version of (CA^2):

$$(\mathrm{ACA}^2)\quad \begin{cases} x_{n+1} \in \varepsilon_n\text{-argmin}\{f(\xi) + \beta_n c(\xi, y_n) + \alpha_n h(x_n, \xi) :\ \xi \in \mathcal{X}\}, \\ y_{n+1} \in \varepsilon_n\text{-argmin}\{g(\eta) + \mu_n c(x_{n+1}, \eta) + \nu_n k(y_n, \eta) : \eta \in \mathcal{Y}\}. \end{cases}$$

In order to describe the asymptotic behavior of the sequences generated by this algorithm, let us define the following notion of equilibrium.

DEFINITION 1.1. *Given nonnnegative constants α, β, μ, ν, an inertial Nash equilibrium, in short INE(α,β,μ,ν), is defined as a couple $(\overline{x}, \overline{y}) \in \mathcal{X} \times \mathcal{Y}$ such that*

- $\overline{x} \in \arg\min\{f(\xi) + \beta c(\xi, \overline{y}) + \alpha h(\overline{x}, \xi) : \xi \in \mathcal{X}\}$,
- $\overline{y} \in \arg\min\{g(\eta) + \mu c(\overline{x}, \eta) + \nu k(\overline{y}, \eta) : \eta \in \mathcal{Y}\}$.

Remark 1. When $\alpha = \nu = 0$, one recovers the usual notion of Nash equilibrium:

- $\overline{x} \in \arg\min\{F(\xi, \overline{y}) : \xi \in \mathcal{X}\}$,
- $\overline{y} \in \arg\min\{G(\overline{x}, \eta) : \eta \in \mathcal{Y}\}$,

where $F(\xi, \eta) = f(\xi) + \beta c(\xi, \eta)$, $G(\xi, \eta) = g(\eta) + \mu c(\xi, \eta)$.

Remark 2. Suppose the sequences α_n, β_n, μ_n, ν_n converge and set $\alpha = \lim \alpha_n$, $\beta = \lim \beta_n$, $\mu = \lim \mu_n$, and $\nu = \lim \nu_n$. We shall prove in Proposition 2.2 that the limit $(\overline{x}, \overline{y})$ of any sequence (x_n, y_n) generated by algorithm (CA^2) is an inertial Nash equilibrium INE(α,β,μ,ν). The terminology Nash equilibrium will be justified in section 1.4.

Algorithm (CA^2) bears a straight relationship with two classical topics:

(i) proximal algorithms and dynamical systems in optimization,

(ii) best response dynamics in game theory.

Let us make some of these links precise; this will help us motivate and introduce some important aspects of this paper (interpretation of the equilibria, and convergence properties of the algorithm).

1.3. Links with proximal algorithms and dynamical optimization. Dynamical systems provide a rich and unifying approach to the analysis of a number of iterative proximal algorithms in optimization (e.g., [4, 5, 11, 3, 6, 20, 25]).

Let $\varphi : H \mapsto \mathbb{R} \cup \{+\infty\}$ be a closed convex proper function on the Hilbert space H. Classical proximal algorithms for convex optimization

$$x_{n+1} = \mathrm{prox}_{\varphi} x_n = (I + \partial\varphi)^{-1} x_n$$

bear a close relationship with the steepest descent dynamical system

$$\dot{x}(t) + \partial\varphi(x(t)) \ni 0.$$

Indeed, they can be viewed as implicit discretizations of this differential inclusion. In the same spirit, if we recall that alternating proximal minimization algorithms minimize (in the convex setting) the bivariate function

$$L(x, y) = f(x) + g(y) + \frac{1}{2} \parallel x - y \parallel^2,$$

then it is natural to associate the steepest descent continuous dynamical system to L in $H \times H$. We obtain the system of coupled differential inclusions

$$\begin{cases} \dot{x}(t) + \partial f(x(t)) + (x(t) - y(t)) \ni 0, \\ \dot{y}(t) + \partial g(y(t)) + (y(t) - x(t)) \ni 0. \end{cases}$$

Noticing that L is closed convex on $H \times H$ we know, by Bruck's theorem [19], that $(x(t), y(t))$ weakly converges to a minimum point $(\overline{x}, \overline{y})$ of L. Thus, one can expect similar results by discretizing this system in an alternating and implicit way. Following the chain computation $y_n \to x_{n+1} \to y_{n+1} \to x_{n+2}$, one obtains

$$\begin{cases} \frac{1}{\lambda_n}(x_{n+1} - x_n) + \partial f(x_{n+1}) + (x_{n+1} - y_n) \ni 0, \\ \frac{1}{\lambda_n}(y_{n+1} - y_n) + \partial g(y_{n+1}) + (y_{n+1} - x_{n+1}) \ni 0. \end{cases}$$

Equivalently, one gets the following (CA^2) algorithm:

$$\begin{cases} x_{n+1} \in \operatorname{argmin}\{f(\xi) + \frac{1}{2\lambda_n} \parallel \xi - x_n \parallel^2 + \frac{1}{2} \parallel \xi - y_n \parallel^2 :\ \xi \in H\}, \\ y_{n+1} \in \operatorname{argmin}\{g(\eta) + \frac{1}{2\lambda_n} \parallel \eta - y_n \parallel^2 + \frac{1}{2} \parallel \eta - x_{n+1} \parallel^2 :\ \eta \in H\}. \end{cases}$$

Note the novelty consisting in introducing the additional terms $\frac{1}{2\lambda_n} \parallel \xi - x_n \parallel^2$ and $\frac{1}{2\lambda_n} \parallel \eta - y_n \parallel^2$ in these variational formulations. Their contributions $\parallel x_{n+1} - x_n \parallel^2$ and $\parallel y_{n+1} - y_n \parallel^2$ will vanish asymptotically. But they are important for the analysis of these algorithms, because they make the tools of dissipative dynamical systems available. Indeed, they make the bifunction L a *strict* Lyapunov function (without these terms $L(x_n, y_n)$ is only nonincreasing); see section 2 for precise statements. Moreover, these dissipative properties allow us to introduce and study these algorithms in a general setting.

1.4. Links with decision sciences and game theory: Inertial Nash equilibration processes. We do not aim at modeling; we just want to stress some relationships of the algorithm (CA^2) with decision sciences and justify the introduction of costs to move. Let us examplify it in the context of noncooperative dynamical game theory with real world interacting players.

"Inertia free" alternating games. Consider two interrelated players 1 and 2 departing from strict individualism to take each other's decision into account via a coupling function. Their static payoffs are made of two components: an individual payoff coming from their own action (decision, strategy, performance, etc., can be used as well) and a common payoff coming from their joint actions. Let $\mathcal{X}$ and $\mathcal{Y}$ be the strategy sets of players 1 and 2, with $\xi \in \mathcal{X}$ and $\eta \in \mathcal{Y}$ their respective current actions. Their static loss functions are

$$\begin{aligned} F &: (\xi, \eta) \in \mathcal{X} \times \mathcal{Y} \mapsto F(\xi, \eta) = f(\xi) + \beta c(\xi, \eta), \\ G &: (\xi, \eta) \in \mathcal{X} \times \mathcal{Y} \mapsto G(\xi, \eta) = g(\eta) + \mu c(\xi, \eta). \end{aligned}$$

The coupling term defines the more or less conflictual characteristic of the game, i.e., the nature of the interdependence between players. Coefficients $\beta > 0$ and $\mu > 0$ represent how much each player benefits from the joint payoff.

Inertial nonautonomous Nash equilibration processes. (i) First consider inertia aspects. We add to this classical static normal form game some costs to move for each player. Player 1 must pay the cost $h(x, \xi)$ to move from action $x \in \mathcal{X}$ to a new action $\xi \in \mathcal{X}$ and player 2 must pay the cost $k(y, \eta)$ to move from action $y \in \mathcal{Y}$ to a new action $\eta \in \mathcal{Y}$.

The general idea is that, in real life, changing, improving the gain, and the quality of actions has a cost. "Costs to move" covers various physical, physiological, psychological, and cognitive aspects. They reflect the bounded rationality and behavioral features of decision processes in real life (see Kahneman [27], Camerer and Loewenstein [21]; Simon [37] for the concept of deliberation costs; Attouch and Soubeyran

[9, 10] for the precise concept of costs to change; and Attouch, Buttazzo, and Michaille [7, section 3.4.2] for a dynamical cognitive approach of Ekeland's variational principle due to Attouch and Soubeyran). Here, these costs mainly describe an anchoring effect. Agents have a (local) vision of their environment which depends on their current actions. Each action is anchored to the preceding one, which means that the perception the agents have of the quality of their subsequent actions depends on the current ones. In economics and management, one may think of actions as routines, ways of doing, while costs to change reflect the difficulty of quitting a routine or entering another one or reacting quickly (reactivity costs). In our situation, suppose that the current action of the two players is $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Suppose also that only player 1 can choose a new action (i.e., player 1 chooses a new action ξ while player 2 "stays" at y); then the inertial payoff of player 1 is

$$F(\xi, y) + \alpha h(x, \xi) = f(\xi) + \beta c(\xi, y) + \alpha h(x, \xi).$$

The second member of this expression is the sum of three costs: a cost to be far from the objective (frustration), a cost to be far from (or close to) the action of the other agent (coupling), and a cost to be far from the preceding action (anchoring or inertial effect). The coefficient α before the cost $h(x, \xi)$ usually reflects some dynamical cognitive features of player 1 (speed, reactivity, learning ability, etc.).

Symmetrically, suppose now that only player 2 can choose a new action (i.e., player 2 chooses a new action η while player 1 "stays" at x); then the inertial payoff of player 2 is

$$G(x, \eta) + \nu k(y, \eta) = g(\eta) + \mu c(x, \eta) + \nu k(y, \eta)$$

with the ν coefficient reflecting some dynamical cognitive features of player 2. The timing of the game follows an asynchronous dynamic where players move in alternation.

(ii) Then consider nonautonomous aspects. Suppose that both the weights $\beta_n > 0$ and $\mu_n > 0$ attached to the joint payoff and the "anchoring" (also called "inertial") coefficients $\alpha_n > 0$ and $\nu_n > 0$ vary from period to period for each player. Suppose that the current action of the two players is $(x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$. Suppose that only player 1 can choose a new action ξ; then the inertial payoff of player 1 is

$$F_n(\xi, y_n) + \alpha_n h(x_n, \xi) = f(\xi) + \beta_n c(\xi, y_n) + \alpha_n h(x_n, \xi).$$

Symmetrically, suppose now that only player 2 can choose a new action; then the inertial payoff of player 2 is

$$G_n(x_n, \eta) + \nu_n k(y_n, \eta) = g(\eta) + \mu_n c(x_n, \eta) + \nu_n k(y_n, \eta).$$

These expressions describe the nonautonomous inertial payoffs of the two agents. Note that they depend both on the current period (time dependence) and the current position of the agents. Taking account of the fact that agents optimize their payoffs, we obtain the description of the alternating dynamic as (CA^2).

1.5. Examples. Algorithm (CA^2) covers, but is more general than, former algorithms to be found in the literature.

Suppose that f and g are proper convex lower semicontinuous functions defined on a Hilbert space, that the costs to move h and k vanish, and that the coupling c is a quadratic dissimilarity cost, namely $c(x, y) = \frac{1}{2} \parallel x - y \parallel^2$.

Then writing out the optimality conditions for algorithm (CA^2) yields

$$\begin{cases} 0 \in \partial f(x_{n+1}) + \beta_n(x_{n+1} - y_n), \\ 0 \in \partial g(y_{n+1}) + \mu_n(y_{n+1} - x_{n+1}), \end{cases} \quad \text{equivalently} \quad \begin{cases} x_{n+1} = (I + \frac{1}{\beta_n}\partial f)^{-1} y_n, \\ y_{n+1} = (I + \frac{1}{\mu_n}\partial g)^{-1} x_{n+1}, \end{cases}$$

where ∂f and ∂g are the subgradient operators of f and g in the sense of convex analysis [32, 36, 12]. If β_n and μ_n have a common constant value β, then this is exactly Acker and Prestel's alternating process, which is introduced in [1] to minimize the function $f(x) + g(y) + \frac{\beta}{2} \parallel x - y \parallel^2$ and which possesses interesting convergence properties. If β_n and μ_n coincide, and if the sequence $(1/\beta_n)$ belongs to $l^2(\mathbb{N}) \setminus l^1(\mathbb{N})$ (which implies $\beta_n \to +\infty$), then this is Passty's scheme for minimizing $f + g$ which possesses interesting properties too; see [35].

It may be useful to illustrate the behavior of algorithm (CA^2) with an example providing a simple geometric interpretation.

Suppose $\mathcal{X}$ and $\mathcal{Y}$ coincide with the same Hilbert space $\mathcal{H}$ endowed with the norm $\parallel \ \parallel$, $\mathcal{X} = \mathcal{Y} = \mathcal{H}$. Let F be a nonvoid closed convex set of $\mathcal{H}$, and define $f(x) = \frac{1}{2}d^2(x, F)$, where $d(x, F)$ denotes the distance from x to F; likewise define $g(y) = \frac{1}{2}d^2(y, G)$, where G is another nonvoid closed convex set in $\mathcal{H}$. Starting from an initial state (x_0, y_0) in $\mathcal{H} \times \mathcal{H}$, the agents seek to improve their decisions x and y subject to quadratic costs (both coupling and inertial terms) and are supposed to implement algorithm (CA^2):

$$\begin{cases} x_{n+1} = \operatorname{argmin}\{ \frac{1}{2}d^2(\xi, F) + \frac{1}{2}\alpha \parallel \xi - x_n \parallel^2 + \frac{1}{2}\beta \parallel \xi - y_n \parallel^2 \colon \xi \in \mathcal{H} \}, \\ y_{n+1} = \operatorname{argmin}\{ \frac{1}{2}d^2(\eta, G) + \frac{1}{2}\mu \parallel \eta - x_{n+1} \parallel^2 + \frac{1}{2}\nu \parallel \eta - y_n \parallel^2 \colon \eta \in \mathcal{H} \}, \end{cases}$$

where α, β, μ, ν are nonnegative constants (owing to the strict convexity and the coerciveness of the functions to be minimized the sequence (x_n, y_n) is uniquely defined). Due to the simple form of the problem, x_{n+1} and y_{n+1} may be explicitly computed; just write the optimality conditions for x_{n+1} and y_{n+1},

$$\begin{cases} x_{n+1} - P_F x_{n+1} + \alpha(x_{n+1} - x_n) + \beta(x_{n+1} - y_n) = 0, \\ y_{n+1} - P_G y_{n+1} + \mu(y_{n+1} - x_{n+1}) + \nu(y_{n+1} - y_n) = 0. \end{cases}$$

Here, P_F and P_G are the projection operators onto the convex sets F and G; for $\partial(1/2)d^2(x, F) = x - P_F(x)$ in general Hilbert space, see [32, p. 286] or [18, p. 46, example 2.8.2]. Set $u_{n+1} = \frac{1}{\alpha+\beta}(\alpha x_n + \beta y_n)$ and $v_{n+1} = \frac{1}{\mu+\nu}(\mu x_{n+1} + \nu y_n)$. The system above reads

$$\begin{cases} x_{n+1} - P_F x_{n+1} + (\alpha + \beta)(x_{n+1} - u_{n+1}) = 0, \\ y_{n+1} - P_G y_{n+1} + (\mu + \nu)(y_{n+1} - v_{n+1}) = 0, \end{cases}$$

which shows that x_{n+1}, $P_F x_{n+1}$, and u_{n+1} are collinear. Hence $P_F x_{n+1} = P_F u_{n+1}$; likewise $P_F y_{n+1} = P_F v_{n+1}$. Finally, x_{n+1} and y_{n+1} are defined by

$$\begin{cases} x_{n+1} = \frac{1}{1+(\alpha+\beta)}(P_F u_{n+1} + (\alpha + \beta)u_{n+1}), \\ y_{n+1} = \frac{1}{1+(\mu+\nu)}(P_G v_{n+1} + (\mu + \nu)v_{n+1}). \end{cases}$$

Observe that x_{n+1}, y_{n+1} and u_{n+1}, v_{n+1} may be defined in quite simple geometric terms: u_{n+1} is the weighted mean of x_n and y_n with weights α and β, and x_{n+1} itself is the weighted mean of $P_F u_{n+1}$ and u_{n+1} with weights 1 and $\alpha + \beta$; likewise v_{n+1} is the weighted mean of x_{n+1} and y_n with weights μ and ν, and y_{n+1} itself is

the weighted mean of $P_G v_{n+1}$ and v_{n+1} with weights 1 and $\mu+\nu$. Figure 1.1, left, displays the first three steps of the algorithm applied to the example with $\mathcal{H}=\mathbb{R}^2$. Figure 1.1, right, displays the particular case where $\alpha=\nu=0$ and $\beta=\mu$; this is Acker and Prestel's alternating minimization process, and the sequences x_n and y_n are then known to converge to (x_∞, y_∞), a minimizer of $f(x)+g(y)+\frac{\beta}{2}\parallel x-y\parallel^2$.

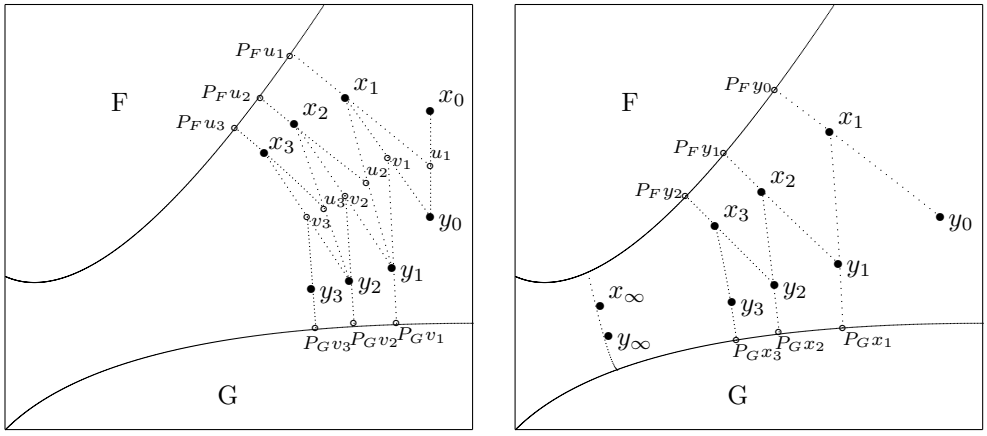


FIG. 1.1. *Illustration of algorithm (CA²) (left); Acker and Prestel's process (right).*

A limit case of the above example occurs when f and g are the respective indicator functions of the sets F and G (that is, $f(x)=0$ if $x\in F$ and $f(x)=+\infty$ if $x\notin F$, and g is similarly defined). The algorithm (CA²) then takes the form

$$\begin{cases} x_{n+1}\in \operatorname{argmin}\{\frac{1}{2}\alpha\parallel \xi-x_n\parallel^2+\frac{1}{2}\beta\parallel \xi-y_n\parallel^2 : \xi\in F\},\\ y_{n+1}\in \operatorname{argmin}\{\frac{1}{2}\mu\parallel \eta-x_{n+1}\parallel^2+\frac{1}{2}\nu\parallel \eta-y_n\parallel^2 : \eta\in G\}. \end{cases}$$

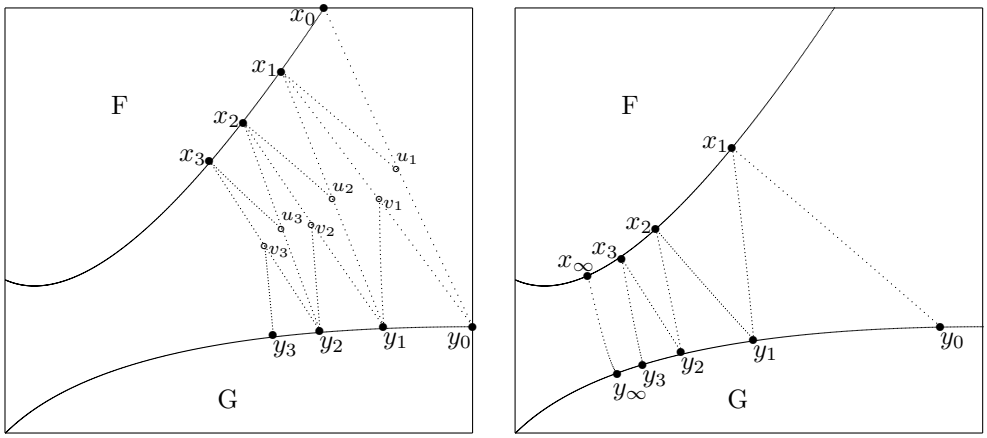


FIG. 1.2. *Illustration of algorithm (CA²) (left); alternating projections (right).*

Writing the optimality conditions for x_{n+1} and y_{n+1} yields

$$\begin{cases} 0\in N_F(x_{n+1})+\alpha(x_{n+1}-x_n)+\beta(x_{n+1}-y_n)=0,\\ 0\in N_G(y_{n+1})+\mu(y_{n+1}-x_{n+1})+\nu(y_{n+1}-y_n)=0, \end{cases}$$

where $N_F(x_{n+1})$ is the outward normal cone to F at point x_{n+1} and $N_G(y_{n+1})$ is the outward normal cone to G at point y_{n+1}. Let us retain the same notation as before for u_{n+1} and v_{n+1}; the above system reads

$$\begin{cases} 0 \in N_F(x_{n+1}) + (x_{n+1} - u_{n+1}), & \text{equivalently } x_{n+1} = P_F u_{n+1}, \\ 0 \in N_G(y_{n+1}) + (y_{n+1} - v_{n+1}), & \text{equivalently } y_{n+1} = P_G v_{n+1}. \end{cases}$$

If $\alpha = \nu = 0$, algorithm (CA^2) reduces to $x_{n+1} = P_F y_n$, $y_{n+1} = P_G x_{n+1}$. This is the celebrated alternating projection algorithm used to find a point in the intersection of the sets, or if the latter is void, to find a couple of points realizing the distance between the two sets; see von Neumann [33], Halperin [24], Bregman [17], and Bauschke, Borwein, and Lewis [14] for further references. Figure 1.2 illustrates this example with $\mathcal{H} = \mathbb{R}^2$.

2. General dissipative and convergence properties of the alternating algorithm (CA^2). Let us fix the general mathematical setting and properties of algorithm (CA^2). In the next two sections, we shall make these results more precise by specifying the type of inertial costs to move which is considered: high local costs to move (section 3) or low local costs to move (section 4). Let us recall that algorithm (CA^2), also called the inertial Nash equilibration process (see section 1.4), is defined by $(x_n, y_n) \to (x_{n+1}, y_n) \to (x_{n+1}, y_{n+1})$ as follows:

$$\begin{cases} x_{n+1} = \operatorname{argmin}\{f(\xi) + \beta_n c(\xi, y_n) + \alpha_n h(x_n, \xi) : \ \xi \in \mathcal{X}\}, \\ y_{n+1} = \operatorname{argmin}\{g(\eta) + \mu_n c(x_{n+1}, \eta) + \nu_n k(y_n, \eta) : \ \eta \in \mathcal{Y}\}. \end{cases}$$

2.1. Convergence to an inertial Nash equilibrium. First let us specify the general topological assumptions in this section.

$\mathcal{H}11$. $\mathcal{X}$ and $\mathcal{Y}$ are topological spaces;

$\mathcal{H}12$. $f : \mathcal{X} \mapsto \mathbb{R} \cup \{+\infty\}$ and $g : \mathcal{Y} \mapsto \mathbb{R} \cup \{+\infty\}$ are proper, bounded below, lower semicontinuous functions;

$\mathcal{H}13$. $h : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}^+$, $k : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}^+$, and $c : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}^+$ are lower semicontinuous and separately continuous;

$\mathcal{H}14$. α_n, β_n, μ_n, and ν_n are nonnegative sequences and $\beta_n \to \beta_\infty$, $\mu_n \to \mu_\infty$.

If algorithm (CA^2) generates a sequence (x_n, y_n) that furthermore converges, then, under some extra assumptions, we are going to prove that the limit is an inertial Nash equilibrium.

PROPOSITION 2.1. *In addition to $\mathcal{H}11$–$\mathcal{H}14$ suppose that $\alpha_n \to \alpha_\infty$ and $\nu_n \to \nu_\infty$. Then the limit $(\overline{x}, \overline{y})$ of any sequence (x_n, y_n) generated by algorithm (CA^2) is an inertial Nash equilibrium INE(α_∞, β_∞, μ_∞, ν_∞).*

Proof. Algorithm (CA^2) reads

$$\begin{aligned} (2.1) \qquad & f(x_{n+1}) + \alpha_n h(x_n, x_{n+1}) + \beta_n c(x_{n+1}, y_n) \\ & \qquad \le f(\xi) + \alpha_n h(x_n, \xi) + \beta_n c(\xi, y_n) \ \forall \xi \in \mathcal{X}, \\ (2.2) \qquad & g(y_{n+1}) + \mu_n c(x_{n+1}, y_{n+1}) + \nu_n k(y_n, y_{n+1}) \\ & \qquad \le g(\eta) + \mu_n c(x_{n+1}, \eta) + \nu_n k(y_n, \eta) \ \forall \eta \in \mathcal{Y}. \end{aligned}$$

The right-hand member of (2.1) converges to $f(\xi) + \alpha_\infty h(\overline{x}, \xi) + \beta_\infty c(\xi, \overline{y})$. Taking the lower limit of the left-hand member of (2.1) gives

$$\begin{aligned} (2.3) \qquad & \underline{\lim}\{f(x_{n+1}) + \alpha_n h(x_n, x_{n+1}) + \beta_n c(x_{n+1}, y_n)\} \\ & \qquad \ge \underline{\lim} f(x_{n+1}) + \underline{\lim} \alpha_n h(x_n, x_{n+1}) + \underline{\lim} \beta_n c(x_{n+1}, y_n) \\ & \qquad \ge f(\overline{x}) + \alpha_\infty h(\overline{x}, \overline{x}) + \beta_\infty c(\overline{x}, \overline{y}), \end{aligned}$$

which yields the first condition for $(\overline{x},\overline{y})$ to be an inertial Nash equilibrium. The second condition is proved likewise from (2.2). □

The proof of the following proposition runs along the same lines.

PROPOSITION 2.2. *In addition to* $\mathcal{H}11$–$\mathcal{H}14$, *suppose*

(i) $h(\xi,\xi)=0, \forall \xi \in \mathcal{X}$, $k(\eta,\eta)=0, \forall \eta \in \mathcal{Y}$;

(ii) $\overline{\alpha}=\overline{\lim}\alpha_n$ *and* $\overline{\nu}=\overline{\lim}\nu_n$ *are finite.*

Then the limit $(\overline{x},\overline{y})$ *of any sequence* (x_n,y_n) *generated by algorithm* (CA^2) *is an inertial Nash equilibrium INE*$(\alpha,\ \beta_\infty,\ \mu_\infty,\ \nu)$, *where* α *is any number greater than or equal to* $\overline{\alpha}$, *and* ν *is any number greater than or equal to* $\overline{\nu}$.

2.2. General dissipative properties of the alternating algorithm (CA^2). The assumptions in this section are

$\mathcal{H}21$. $\mathcal{X}$ and $\mathcal{Y}$ are abstract spaces;

$\mathcal{H}22$. $f:\mathcal{X}\mapsto\mathbb{R}\cup\{+\infty\}$ and $g:\mathcal{Y}\mapsto\mathbb{R}\cup\{+\infty\}$ are proper, bounded below;

$\mathcal{H}23$. $h:\mathcal{X}\times\mathcal{X}\mapsto\mathbb{R}^+$, $k:\mathcal{Y}\times\mathcal{Y}\mapsto\mathbb{R}^+$ and $c:\mathcal{X}\times\mathcal{Y}\mapsto\mathbb{R}^+$ are nonnegative functions satisfying $h(x,x)=k(y,y)=0\,\forall\, x\in\mathcal{X},\ y\in\mathcal{Y}$;

$\mathcal{H}24$. $\beta_n>0,\ \mu_n>0,\ \alpha_n\geq 0,\ \nu_n\geq 0$.

LEMMA 2.3. *Assume* $\mathcal{H}21$–$\mathcal{H}24$. *Then, for any sequence* (x_n,y_n) *generated by* (CA^2), *one has*

$$
\begin{aligned}
(2.4)\qquad &[\mu_n f(x_{n+1})+\beta_n g(y_{n+1})+\beta_n\mu_n c(x_{n+1},y_{n+1})]\\
&\qquad+[\mu_n\alpha_n h(x_n,x_{n+1})+\beta_n\nu_n k(y_n,y_{n+1})]\\
&\qquad\leq[\mu_n f(x_n)+\beta_n g(y_n)+\beta_n\mu_n c(x_n,y_n)].
\end{aligned}
$$

Proof. Set $\xi=x_n$ and $\eta=y_n$ in (2.1) and in (2.2); sequences x_n and y_n verify

$$
\begin{cases}
f(x_{n+1})+\alpha_n h(x_n,x_{n+1})+\beta_n c(x_{n+1},y_n)\leq f(x_n)+\beta_n c(x_n,y_n),\\
g(y_{n+1})+\mu_n c(x_{n+1},y_{n+1})+\nu_n k(y_n,y_{n+1})\leq g(y_n)+\mu_n c(x_{n+1},y_n).
\end{cases}
$$

Multiplying the first inequality by μ_n and the second one by β_n we obtain

$$
\begin{aligned}
&[\mu_n f(x_{n+1})]+[\mu_n\alpha_n h(x_n,x_{n+1})]+[\mu_n\beta_n c(x_{n+1},y_n)]\\
&\qquad\leq[\mu_n f(x_n)+\mu_n\beta_n c(x_n,y_n)]
\end{aligned}
$$

and

$$
\begin{aligned}
&[\beta_n g(y_{n+1})+\beta_n\mu_n c(x_{n+1},y_{n+1})]+[\beta_n\nu_n k(y_n,y_{n+1})]\\
&\qquad\leq[\beta_n g(y_n)]+[\beta_n\mu_n c(x_{n+1},y_n)].
\end{aligned}
$$

Add the two inequalities to get (2.4). □

Depending on various monotonicity assumptions on coefficients β_n and μ_n, inequality (2.4) gives prominence to quantities decreasing along the trajectories, and hence acting as Lyapunov functions. We first consider the particular case where $\beta_n\equiv\mu_n$ before the general case.

THEOREM 2.4. *In addition to assumption* $\mathcal{H}21$–$\mathcal{H}24$ *assume that* β_n *and* μ_n *coincide and that* $\beta_n\equiv\mu_n$ *is a nonincreasing sequence. Then, the following marginal analysis result holds: for any sequence* (x_n,y_n) *generated by* (CA^2), *one has*

(a) $f(x_n)+g(y_n)+\beta_n c(x_n,y_n)$ *is nonincreasing and has a limit as* $n\to\infty$;

(b) $\sum_{n=0}^{+\infty}[\alpha_n h(x_n,x_{n+1})+\nu_n k(y_n,y_{n+1})]<+\infty$.

Proof. Dividing inequality (2.4) by $\beta_n\equiv\mu_n$, we obtain

$$
\begin{aligned}
&[f(x_{n+1})+g(y_{n+1})+\beta_n c(x_{n+1},y_{n+1})]+[\alpha_n h(x_n,x_{n+1})+\nu_n k(y_n,y_{n+1})]\\
&\qquad\leq[f(x_n)+g(y_n)+\beta_n c(x_n,y_n)].
\end{aligned}
$$

Since β_n is nonincreasing, we may replace β_n with β_{n+1} in the left-hand member,

$$[f(x_{n+1}) + g(y_{n+1}) + \beta_{n+1}c(x_{n+1}, y_{n+1})] + [\alpha_n h(x_n, x_{n+1}) + \nu_n k(y_n, y_{n+1})] \\ \leq [f(x_n) + g(y_n) + \beta_n c(x_n, y_n)],$$

which shows that the quantity $f(x_n) + g(y_n) + \beta_n c(x_n, y_n)$ is nonincreasing; as it is nonnegative it is convergent, which proves point (a).

Summing the inequality above from $n = 0$ to N, we obtain

$$\sum_{n=0}^{N} [\alpha_n h(x_n, x_{n+1}) + \nu_n k(y_n, y_{n+1})] \\ \leq [f(x_0) + g(y_0) + \beta_0 c(x_0, y_0)] - [f(x_{N+1}) + g(y_{N+1}) + \beta_{N+1} c(x_{N+1}, y_{N+1})] \\ \leq [f(x_0) + g(y_0) + \beta_0 c(x_0, y_0)] - [\inf_{\mathcal{X}} f + \inf_{\mathcal{Y}} g],$$

which proves point (b). □

The preceding theorem can be extended to monotone (i.e., nonincreasing or nondecreasing) sequences β_n and μ_n. Define δ_n according to the following cases:

- if β_n and μ_n are nonincreasing sequences, then $\delta_n = 1$;
- if β_n and μ_n are nondecreasing sequences, then $\delta_n = \beta_n \mu_n$;
- if β_n is nonincreasing and μ_n is nondecreasing, then $\delta_n = \mu_n$;
- if β_n is nondecreasing and μ_n is nonincreasing, then $\delta_n = \beta_n$.

THEOREM 2.5. *Assume $\mathcal{H}21$–$\mathcal{H}24$ and let (x_n, y_n) be any sequence generated by (CA^2). If β_n and μ_n are monotone sequences, then*

(a) $\frac{\mu_n}{\delta_n}(f(x_n) - \inf_{\mathcal{X}} f) + \frac{\beta_n}{\delta_n}(g(y_n) - \inf_{\mathcal{Y}} g) + \frac{\beta_n \mu_n}{\delta_n} c(x_n, y_n)$ *is nonincreasing and has a limit as $n \to \infty$;*

(b) $\sum_{n=0}^{\infty} \left[\frac{\mu_n \alpha_n}{\delta_n} h(x_n, x_{n+1}) + \frac{\beta_n \nu_n}{\delta_n} k(y_n, y_{n+1})\right] < +\infty$.

Proof. Divide inequality (2.4) in Lemma 2.3 by δ_n, and display positive quantities like $f(x_n) - \inf_{\mathcal{X}} f$ and $g(y_n) - \inf_{\mathcal{Y}} g$. Then

$$\left[\frac{\mu_n}{\delta_n}\left(f(x_{n+1}) - \inf_{\mathcal{X}} f\right) + \frac{\beta_n}{\delta_n}\left(g(y_{n+1}) - \inf_{\mathcal{Y}} g\right) + \frac{\beta_n \mu_n}{\delta_n} c(x_{n+1}, y_{n+1})\right] \\ + \left[\frac{\mu_n \alpha_n}{\delta_n} h(x_n, x_{n+1}) + \frac{\beta_n \nu_n}{\delta_n} k(y_n, y_{n+1})\right] \\ \leq \left[\frac{\mu_n}{\delta_n}\left(f(x_n) - \inf_{\mathcal{X}} f\right) + \frac{\beta_n}{\delta_n}\left(g(y_n) - \inf_{\mathcal{Y}} g\right) + \frac{\beta_n \mu_n}{\delta_n} c(x_n, y_n)\right].$$

Each sequence $\frac{\mu_n}{\delta_n}$, $\frac{\beta_n}{\delta_n}$, and $\frac{\beta_n \mu_n}{\delta_n}$ is nonincreasing; so we have

$$\left[\frac{\mu_{n+1}}{\delta_{n+1}}\left(f(x_{n+1}) - \inf_{\mathcal{X}} f\right) + \frac{\beta_{n+1}}{\delta_{n+1}}\left(g(y_{n+1}) - \inf_{\mathcal{Y}} g\right) + \frac{\beta_{n+1} \mu_{n+1}}{\delta_{n+1}} c(x_{n+1}, y_{n+1})\right] \\ + \left[\frac{\mu_n \alpha_n}{\delta_n} h(x_n, x_{n+1}) + \frac{\beta_n \nu_n}{\delta_n} k(y_n, y_{n+1})\right] \\ \leq \left[\frac{\mu_n}{\delta_n}\left(f(x_n) - \inf_{\mathcal{X}} f\right) + \frac{\beta_n}{\delta_n}\left(g(y_n) - \inf_{\mathcal{Y}} g\right) + \frac{\beta_n \mu_n}{\delta_n} c(x_n, y_n)\right],$$

which completes the proof. □

3. High local costs to move. The following assumptions will be valid throughout this section:

- $\mathcal{H}31$. $\mathcal{X}$ and $\mathcal{Y}$ are complete metric spaces endowed with distances $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$;
- $\mathcal{H}32$. $f : \mathcal{X} \mapsto \mathbb{R} \cup \{+\infty\}$ and $g : \mathcal{Y} \mapsto \mathbb{R} \cup \{+\infty\}$ are proper, bounded below, lower semicontinuous functions;
- $\mathcal{H}33$. $c : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}^+ \cup \{+\infty\}$ is a proper, lower semicontinuous function;
- $\mathcal{H}34$. $h : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}^+$ and $k : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}^+$ are continuous functions satisfying $h(\xi,\xi) = k(\eta,\eta) = 0$ and $h(x,\xi) \geq d_{\mathcal{X}}(x,\xi)$, $k(y,\eta) \geq d_{\mathcal{Y}}(y,\eta)\, \forall\, (x,\xi,y,\eta)$ in $\mathcal{X}^2 \times \mathcal{Y}^2$;
- $\mathcal{H}35$. β_n and μ_n are monotone (i.e., nonincreasing or nondecreasing) converging sequences: $\beta_\infty = \lim_n \beta_n$, $\mu_\infty = \lim_n \mu_n$;
- $\mathcal{H}36$. α_n and ν_n are nonnegative sequences, bounded and bounded away from 0: $0 < \underline{\alpha} = \underline{\lim}\alpha_n \leq \overline{\alpha} = \overline{\lim}\alpha_n < +\infty$, $0 < \underline{\nu} = \underline{\lim}\nu_n \leq \overline{\nu} = \overline{\lim}\nu_n < +\infty$;
- $\mathcal{H}37$. ε_n is a positive sequence such that $\sum_{n=0}^{+\infty} \varepsilon_n < \infty$.

Of importance are the completeness of spaces $\mathcal{X}$ and $\mathcal{Y}$, the inequalities $h \geq d_{\mathcal{X}}$, and $k \geq d_{\mathcal{Y}}$ in $\mathcal{H}34$. We call this last condition *high local costs to move*, in contrast with a condition like $h = d_{\mathcal{X}}^2$, $k = d_{\mathcal{Y}}^2$ which expresses low local costs to move (see section 4).

3.1. Convergence of algorithm (ACA^2) to an inertial Nash equilibrium. Theorem 2.5 can very simply be used to demonstrate the convergence of the sequence (x_n, y_n) when the "local" costs to move are significant; in decision terms it means that the agents are reluctant to switch from their current state to another state because the cost to move is a deterrent even if the next state lies in a neighborhood of the current one. It may be guessed that the agents, in this evolution process, will stop somewhere in the end. The following theorem gives a precise mathematical meaning to this remark.

THEOREM 3.1. *In addition to hypotheses $\mathcal{H}31$–$\mathcal{H}37$ assume that c is separately continuous. Then any sequence generated by algorithm (ACA^2) converges to an inertial Nash equilibrium INE($\overline{\alpha}$,β_∞,μ_∞,$\overline{\nu}$), i.e.,*

$$\begin{cases} \overline{x} \in \mathit{argmin}\{f(\xi) + \overline{\alpha} h(\overline{x},\xi) + \beta_\infty c(\xi,\overline{y}) : \xi \in \mathcal{X}\}, \\ \overline{y} \in \mathit{argmin}\{g(\eta) + \mu_\infty c(\overline{x},\eta) + \overline{\nu} k(\overline{y},\eta) : \eta \in \mathcal{Y}\}. \end{cases}$$

In particular, INE($\overline{\alpha}$,β_∞,μ_∞,$\overline{\nu}$) is nonempty.

Proof. Let (x_n, y_n) be a sequence generated by algorithm (ACA^2). Using the same reasoning as in Theorem 2.5, and with the same definition for δ_n, we obtain

$$\begin{aligned} (3.1)\quad & \left[\frac{\mu_{n+1}}{\delta_{n+1}}\left(f(x_{n+1}) - \inf_{\mathcal{X}} f\right) + \frac{\beta_{n+1}}{\delta_{n+1}}\left(g(y_{n+1}) - \inf_{\mathcal{Y}} g\right) + \frac{\beta_{n+1}\mu_{n+1}}{\delta_{n+1}} c(x_{n+1}, y_{n+1})\right] \\ & + \left[\frac{\mu_n \alpha_n}{\delta_n} h(x_n, x_{n+1}) + \frac{\beta_n \nu_n}{\delta_n} k(y_n, y_{n+1})\right] \\ & \leq \left[\frac{\mu_n}{\delta_n}\left(f(x_n) - \inf_{\mathcal{X}} f\right) + \frac{\beta_n}{\delta_n}\left(g(y_n) - \inf_{\mathcal{Y}} g\right) + \frac{\beta_n \mu_n}{\delta_n} c(x_n, y_n)\right] + \left(\frac{\mu_n}{\delta_n} + \frac{\beta_n}{\delta_n}\right)\varepsilon_n. \end{aligned}$$

In view of the definition of δ_n, according to the behavior of β_n and μ_n, the sequences μ_n/δ_n and β_n/δ_n are nonincreasing; hence we have

$$\sum_{n=0}^{\infty}\left(\frac{\mu_n}{\delta_n} + \frac{\beta_n}{\delta_n}\right)\varepsilon_n \leq \left(\frac{\mu_0}{\delta_0} + \frac{\beta_0}{\delta_0}\right)\sum_{n=0}^{\infty} \varepsilon_n. < +\infty.$$

Adding the inequalities (3.1), as in the proof of Theorem 2.5, yields the convergence of the series

$$\sum_{n=0}^{\infty} \frac{\mu_n}{\delta_n}\alpha_n h(x_n, x_{n+1}) + \frac{\beta_n}{\delta_n}\nu_n k(y_n, y_{n+1}) < +\infty.$$

Let δ_∞ be the limit of sequence δ_n (it does converge). In view of $\mathcal{H}36$ there exists some $N \in \mathbb{N}$ such that

$$\frac{\mu_n}{\delta_n}\alpha_n > \frac{1}{2}\frac{\mu_\infty}{\delta_\infty}\underline{\alpha} > 0, \quad \frac{\beta_n}{\delta_n}\nu_n > \frac{1}{2}\frac{\beta_\infty}{\delta_\infty}\underline{\nu} > 0 \ \ \forall n \geq N.$$

Hence

$$\sum_{n=0}^{\infty} h(x_n, x_{n+1}) + k(y_n, y_{n+1}) < +\infty$$

and finally

$$\sum_{n=0}^{\infty} d_{\mathcal{X}}(x_n, x_{n+1}) + d_{\mathcal{Y}}(y_n, y_{n+1}) < +\infty.$$

The sequences (x_n) and (y_n) are thus Cauchy sequences in the complete metric spaces $\mathcal{X}$ and $\mathcal{Y}$. Let $\overline{x}$ and $\overline{y}$ be their limits.

Recall that, by the definition of x_{n+1}, we have, for every $\xi \in \mathcal{X}$,

$$f(x_{n+1}) + \alpha_n h(x_n, x_{n+1}) + \beta_n c(x_{n+1}, y_{n+1}) \leq f(\xi) + \alpha_n h(x_n, \xi) + \beta_n c(\xi, y_n).$$

Taking the lower limit of the left-hand side yields

$$\begin{aligned}\liminf\{f(x_{n+1}) + \alpha_n h(x_n, x_{n+1}) + \beta_n c(x_{n+1}, y_n)\} \\ \geq \liminf f(x_{n+1}) + \liminf \alpha_n h(x_n, x_{n+1}) + \liminf \beta_n c(x_{n+1}, y_n) \\ \geq f(\overline{x}) + \beta_\infty c(\overline{x}, \overline{y}).\end{aligned}$$

And taking the upper limit of the right-hand side yields

$$\limsup\{f(\xi) + \alpha_n h(x_n, \xi) + \beta_n c(\xi, y_n)\} \leq f(\xi) + \overline{\alpha} h(\overline{x}, \xi) + \beta_\infty c(\xi, \overline{y}).$$

Hence

$$f(\overline{x}) + \beta_\infty c(\overline{x}, \overline{y}) \leq f(\xi) + \overline{\alpha} h(\overline{x}, \xi) + \beta_\infty c(\xi, \overline{y}).$$

Likewise we can prove

$$g(\overline{y}) + \mu_\infty c(\overline{x}, \overline{y}) \leq g(\eta) + \overline{\nu} k(\overline{y}, \eta) + \mu_\infty c(\overline{x}, \eta) \ \forall \eta \in \mathcal{Y}.$$

And that shows that $(\overline{x}, \overline{y})$ is an inertial Nash equilibrium INE($\overline{\alpha}$, β_∞, μ_∞, $\overline{\nu}$). ∎

Let us come back to the decision theory point of view. If the costs to move h and k are locally low, then we can imagine that the agents are inclined to trying to improve his lot. The inequalities $h(x, \xi) \geq d^2_{\mathcal{X}}(x, \xi)$ and $k(y, \eta) \geq d^2_{\mathcal{Y}}(y, \eta)$ may account for low costs to move; they are locally (i.e., for small distances) less stringent than $h(x, \xi) \geq d_{\mathcal{X}}(x, \xi)$ and $k(y, \eta) \geq d_{\mathcal{Y}}(y, \eta)$. Then the convergence of the series $(h(x_n, x_{n+1}))$ and $(k(y_n, y_{n+1}))$ only yields $\Sigma_{n=0}^{+\infty} d^2_{\mathcal{X}}(x_n, x_{n+1}) < +\infty$ and $\Sigma_{n=0}^{+\infty} d^2_{\mathcal{Y}}(y_n, y_{n+1}) < +\infty$ which is not enough to guarantee the convergence of the sequence (x_n, y_n). A parallel may be drawn with the friction phenomenon in mechanics: dry friction, with a potential proportional to the modulus of the velocity, usually leads a system to rest within a finite time interval, while viscous friction, with a potential proportional to the square of the velocity modulus, usually leads to rest only asymptotically and may even fail to reach an equilibrium state (see [2, 8]).

3.2. Links with proximal algorithms. For (x_0, y_0) given in $\mathcal{X} \times \mathcal{Y}$, consider the following algorithm:

$$(3.2)\qquad (x_{n+1}, y_{n+1}) \in \operatorname{argmin}\{\ \mu_n f(\xi) + \beta_n g(\eta) + \mu_n\beta_n c(\xi,\eta) + \mu_n\alpha_n h(x_n,\xi) + \beta_n\nu_n k(y_n,\eta),\ (\xi,\eta) \in \mathcal{X}\times\mathcal{Y}\}.$$

It is not difficult to realize that its alternating version is exactly algorithm (CA^2):

- first fix $\eta = y_n$ and minimize with respect to ξ, which yields some x_{n+1} satisfying

$$x_{n+1} = \operatorname{argmin}\{f(\xi) + \alpha_n h(x_n,\xi) + \beta_n c(\xi, y_n),\ \xi \in \mathcal{X}\},$$

- then fix $\xi = x_{n+1}$ and minimize with respect to η, which yields some y_{n+1} satisfying

$$y_{n+1} = \operatorname{argmin}\{g(\eta) + \mu_n c(x_{n+1},\eta) + \nu_n k(y_n,\eta),\ \eta \in \mathcal{Y}\}.$$

Now algorithm (3.2) may be termed a proximal algorithm. Indeed, at each step the sum of the function $(\xi,\eta) \mapsto \mu_n f(\xi) + \beta_n g(\eta) + \mu_n\beta_n c(\xi,\eta)$ and of a perturbation term around (x_n, y_n), namely $\mu_n\alpha_n h(x_n,\xi) + \beta_n\nu_n k(y_n,\eta)$, is to be minimized. Besides, if $\beta_n \equiv \mu_n \equiv \beta$, and if $h = d_{\mathcal{X}}$ (or $d^2_{\mathcal{X}}$) and $k = d_{\mathcal{Y}}$ (or $d^2_{\mathcal{Y}}$), then algorithm (3.2) is exactly the classical proximal algorithm applied to the function $(x,y) \mapsto f(x) + g(y) + \beta c(x,y)$.

Thus algorithm (CA^2) appears as the alternating version of a proximal-like algorithm applied to $\mu_n f + \beta_n g + \beta_n\mu_n c$ with h and k acting as distances on $\mathcal{X}$ and $\mathcal{Y}$.

4. Low local costs to move: The convex case with quadratic costs. In this section we assume the following:

- $\mathcal{H}41$. $\mathcal{X}$ and $\mathcal{Y}$ coincide with the same Hilbert space $\mathcal{H} = \mathcal{X} = \mathcal{Y}$, endowed with the inner product $\langle x, y\rangle$ and the norm $\| x \| = \langle x, x\rangle^{\frac{1}{2}}$;
- $\mathcal{H}42$. functions $f, g : \mathcal{H} \mapsto \mathbb{R} \cup \{+\infty\}$ are convex, lower semicontinuous and proper, with $\inf_{\mathcal{H}} f > -\infty$, $\inf_{\mathcal{H}} g > -\infty$;
- $\mathcal{H}43$. costs to change and dissimilarity costs are quadratic: $h(\xi,\eta) = k(\xi,\eta) = c(\xi,\eta) = c(\xi,\eta) = \frac{1}{2} \| \xi - \eta \|^2\ \forall (\xi,\eta) \in \mathcal{H}^2$;
- $\mathcal{H}44$. $\alpha_n \geq 0$, $\nu_n \geq 0$;
- $\mathcal{H}45$. β_n and μ_n are constant and positive: $\beta_n \equiv \beta > 0$, $\mu_n \equiv \mu > 0$;
- $\mathcal{H}46$. the set of inertial Nash equilibria is nonvoid; namely there exists some $(x,y) \in \mathcal{H}^2$ that is an INE(0,β,μ,0), i.e.,

$$\begin{cases} x \in \operatorname{argmin}\{f(\xi) + \dfrac{\beta}{2} \| \xi - y \|^2 \colon \xi \in \mathcal{H}\}, \\ y \in \operatorname{argmin}\{g(\eta) + \dfrac{\mu}{2} \| \eta - x \|^2 \colon \eta \in \mathcal{H}\}; \end{cases}$$

- $\mathcal{H}47$. $\sum_{n=0}^{+\infty} (\alpha_{n+1} - \alpha_n)^+ < +\infty$, $\sum_{n=0}^{+\infty} (\nu_{n+1} - \nu_n)^+ < +\infty$, and consequently the sequences α_n, ν_n are convergent: $\alpha_n \to \alpha$, $\nu_n \to \nu$.

The quadratic cost hypothesis $\mathcal{H}43$ means that the sequence (x_n, y_n) is generated, and uniquely defined indeed, by the following algorithm:

$$(\mathrm{CA}^2)\quad \begin{cases} x_{n+1} = \operatorname{argmin}\{f(\xi) + \dfrac{\alpha_n}{2} \| \xi - x_n \|^2 + \dfrac{\beta}{2} \| \xi - y_n \|^2 \colon \xi \in \mathcal{H}\}, \\ y_{n+1} = \operatorname{argmin}\{g(\eta) + \dfrac{\mu}{2} \| \eta - x_{n+1} \|^2 + \dfrac{\nu_n}{2} \| \eta - y_n \|^2 \colon \eta \in \mathcal{H}\}. \end{cases}$$

Let us define the function $L_{\beta,\mu} : (\xi,\eta) \in \mathcal{H}^2 \mapsto L_{\beta,\mu}(\xi,\eta) = \mu f(\xi) + \beta g(\eta) + \frac{1}{2}\beta\mu \parallel \xi - \eta \parallel^2$. The purpose of this section is to prove the following theorem.

THEOREM 4.1. *Assume* $\mathcal{H}41$–$\mathcal{H}47$. *Then*

(a) *the sequence* (x_n, y_n) *generated by (CA*^{2}*) weakly converges to some* $(\overline{x},\overline{y})$, *which is an inertial Nash equilibrium INE(*0,β,μ,0*);*

(b) (x_n, y_n) *is a minimizing sequence for the function* $L_{\beta,\mu}$ *and* $(\overline{x},\overline{y})$ *is a minimum point of* $L_{\beta,\mu}$*;*

(c) $x_n - y_n \to \overline{x} - \overline{y}$ *relative to the norm topology in* $\mathcal{H}$, $f(x_n) \to f(\overline{x})$, $g(y_n) \to g(\overline{y})$.

Remark. The convergence asserted in point (a) may fail to be strong; see [26, 31].

The proof of the theorem requires some preparation.

In view of the convexity of f and g, the sequence (x_n, y_n) is characterized by

$$\begin{cases} 0 \in \partial f(x_{n+1}) + \alpha_n(x_{n+1} - x_n) + \beta(x_{n+1} - y_n), \\ 0 \in \partial g(y_{n+1}) + \mu(y_{n+1} - x_{n+1}) + \nu_n(y_{n+1} - y_n), \end{cases}$$

where ∂f and ∂g are the subdifferential operators of f and g in the sense of convex analysis. If we introduce the resolvent operators $T_f^n = (I + \frac{1}{\alpha_n+\beta}\partial f)^{-1}$ and $T_g^n = (I + \frac{1}{\mu+\nu_n}\partial g)^{-1}$ we have equivalently

$$\begin{cases} x_{n+1} = T_f^n\left(\dfrac{\alpha_n x_n + \beta y_n}{\alpha_n + \beta}\right), \\ y_{n+1} = T_g^n\left(\dfrac{\mu x_{n+1} + \nu_n y_n}{\mu + \nu_n}\right). \end{cases} \tag{4.1}$$

Recall that resolvent operators are firmly nonexpansive, i.e.,

$$\langle T_f^n x - T_f^n x', x - x'\rangle \geq \parallel T_f^n x - T_f^n x' \parallel^2 \quad \forall x, x' \in \mathcal{H} \tag{4.2}$$

and likewise for T_g^n; see, e.g., [36, Proposition 1(b)].

The inertial Nash equilibrium (x, y), the existence of which is asserted by assumption $\mathcal{H}46$, is also an inertial Nash equilibrium INE(α,β,μ,ν), where α and ν are arbitrary nonnegative numbers

$$\begin{cases} x \in \operatorname{argmin}\{f(\xi) + \frac{\alpha}{2} \parallel \xi - x \parallel^2 + \frac{\beta}{2} \parallel \xi - y \parallel^2 : \xi \in \mathcal{H}\}, \\ y \in \operatorname{argmin}\{g(\eta) + \frac{\mu}{2} \parallel \eta - x \parallel^2 + \frac{\nu}{2} \parallel \eta - y \parallel^2 : \eta \in \mathcal{H}\}. \end{cases}$$

Indeed, the functions $\xi \mapsto f(\xi) + \frac{\beta}{2} \parallel \xi - y \parallel^2$ and $\xi \mapsto \frac{\alpha}{2} \parallel \xi - x \parallel^2$ are minimum at the same point x. Likewise the functions $\eta \mapsto g(\eta) + \frac{\mu}{2} \parallel \eta - x \parallel^2$ and $\eta \mapsto \frac{\nu}{2} \parallel \eta - y \parallel^2$ are minimum at the same point y.

Conversely, if (x, y) is an inertial Nash equilibrium INE(α,β,μ,ν), then it verifies

$$\begin{cases} 0 \in \partial f(x) + \alpha(x - x) + \beta(x - y), \\ 0 \in \partial g(y) + \mu(y - x) + \nu(y - y), \end{cases} \tag{4.3}$$

and these inclusions are also the optimality conditions for (x, y) to be an inertial Nash equilibrium INE(0,β,μ,0). Now, with $\alpha = \alpha_n$ and $\nu = \nu_n$ in (4.3), we see that (x, y) verifies

$$\begin{cases} x = T_f^n\left(\dfrac{\alpha_n x + \beta y}{\alpha_n + \beta}\right), \\ y = T_g^n\left(\dfrac{\mu x + \nu_n y}{\mu + \nu_n}\right). \end{cases} \tag{4.4}$$

LEMMA 4.2. *Assume $\mathcal{H}41$–$\mathcal{H}45$. The set of inertial Nash equilibria INE($0,\beta,\mu,0$) is the set of minimizers of $L_{\beta,\mu}$.*

Proof. The function $L_{\beta,\mu}$ is proper, lower semicontinuous, and convex. Its subgradient set is easily seen to verify

$$\partial L_{\beta,\mu}(\xi,\eta) = \{\mu\partial f(\xi) + \beta\mu(\xi-\eta)\} \times \{\beta\partial g(\eta) + \beta\mu(\eta-\xi)\}. \tag{4.5}$$

And notice that the conditions (4.3) for (x,y) to be an inertial Nash equilibrium are equivalent to $\partial L_{\beta,\mu}(x,y) \ni 0$. □

LEMMA 4.3. *Assume $\mathcal{H}41$–$\mathcal{H}46$. The sequence (x_n,y_n) generated by (CA^2) verifies the following inequalities:*

$$\text{(a)}\quad \| x_{n+1}-x \|^2 \le \frac{1}{\alpha_n+\beta}\{\, \alpha_n \| x_n-x \|^2 + \beta \| y_n-y \|^2 - \alpha_n \| x_{n+1}-x_n \|^2 - \beta \| y_n - x_{n+1} - y + x \|^2 \,\}.$$

$$\begin{aligned}\text{(b)}\quad \| y_{n+1}-y \|^2 \le \frac{1}{(\mu+\nu_n)(\alpha_n+\beta)}\{\, & \mu\alpha_n \| x_n-x \|^2 \\ & + (\mu\beta + \nu_n(\alpha_n+\beta)) \| y_n-y \|^2 - \mu\alpha_n \| x_{n+1}-x_n \|^2 \\ & - \nu_n(\alpha_n+\beta) \| y_{n+1}-y_n \|^2 - \mu\beta \| y_n - x_{n+1} - y + x \|^2 \\ & - \mu(\alpha_n+\beta) \| y_{n+1} - x_{n+1} - y + x \|^2 \,\}.\end{aligned}$$

$$\begin{aligned}\text{(c)}\quad & \alpha_n\mu \| x_{n+1}-x \|^2 + \beta(\mu+\nu_n) \| y_{n+1}-y \|^2 \\ & \le \alpha_n\mu \| x_n-x \|^2 + \beta(\mu+\nu_n) \| y_n-y \|^2 \\ & \quad - \alpha_n\mu \| x_{n+1}-x_n \|^2 - \beta\nu_n \| y_{n+1}-y_n \|^2 \\ & \quad - \beta\mu \| y_n - x_{n+1} - y + x \|^2 - \beta\mu \| y_{n+1} - x_{n+1} - y + x \|^2.\end{aligned}$$

Proof. The first two inequalities entail the third one; it suffices to multiply the first by $\alpha_n\mu$, the second by $\beta(\mu+\nu_n)$, and to add.

The proof of the first two inequalities relies on the firm nonexpansiveness property of the resolvent operators T_f^n and T_g^n (recall (4.2)). From (4.1) and (4.4) we deduce that

$$\| x_{n+1}-x \|^2 \le \frac{1}{\alpha_n+\beta}\{\alpha_n\langle x_{n+1}-x, x_n-x\rangle + \beta\langle x_{n+1}-x, y_n-y\rangle\},$$
$$\| y_{n+1}-y \|^2 \le \frac{1}{\nu_n+\mu}\{\mu\langle y_{n+1}-y, x_{n+1}-x\rangle + \nu_n\langle y_{n+1}-y, y_n-y\rangle\}.$$

Write each inner product as $\langle u,v\rangle = \frac{1}{2}\{\| u \|^2 + \| v \|^2 - \| v-u \|^2\}$ in the inequalities above. We then obtain

$$\begin{aligned}\| x_{n+1}-x \|^2 \le \frac{1}{\alpha_n+\beta}\Big\{ & \frac{\alpha_n}{2} \| x_{n+1}-x \|^2 + \frac{\alpha_n}{2} \| x_n-x \|^2 - \frac{\alpha_n}{2} \| x_{n+1}-x_n \|^2 \\ & + \frac{\beta}{2} \| x_{n+1}-x \|^2 + \frac{\beta}{2} \| y_n-y \|^2 - \frac{\beta}{2} \| y_n - x_{n+1} - y + x \|^2 \Big\}\end{aligned}$$

and

$$\begin{aligned}\| y_{n+1}-y \|^2 \le \frac{1}{\nu_n+\mu}\Big\{ & \frac{\mu}{2} \| x_{n+1}-x \|^2 + \frac{\mu}{2} \| y_{n+1}-y \|^2 \\ & + \frac{\nu_n}{2} \| y_{n+1}-y \|^2 + \frac{\nu_n}{2} \| y_n-y \|^2 \\ & - \frac{\mu}{2} \| y_{n+1} - x_{n+1} - y + x \|^2 - \frac{\nu_n}{2} \| y_{n+1}-y_n \|^2 \Big\}.\end{aligned}$$

Rearranging the terms we obtain

$$\| x_{n+1} - x \|^2 \leq \frac{1}{\alpha_n + \beta} \{ \alpha_n \| x_n - x \|^2 + \beta \| y_n - y \|^2 - \alpha_n \| x_{n+1} - x_n \|^2 - \beta \| y_n - x_{n+1} - y + x \|^2 \}$$

and

$$\| y_{n+1} - y \|^2 \leq \frac{1}{\nu_n + \mu} \{ \mu \| x_{n+1} - x \|^2 + \nu_n \| y_n - y \|^2 - \mu \| y_{n+1} - x_{n+1} - y + x \|^2 - \nu_n \| y_{n+1} - y_n \|^2 \}.$$

Now, in the last inequality, replace $\| x_{n+1} - x \|^2$ by its upper bound given by the last one to achieve the first two inequalities asserted by the lemma. □

PROPOSITION 4.4. *Assume $\mathcal{H}41$–$\mathcal{H}46$. The sequence (x_n, y_n) generated by (CA^2) is bounded.*

Proof. As a straightforward consequence of Lemma 4.3 we have

$$\begin{aligned} \| x_{n+1} - x \|^2 &\leq \frac{1}{\alpha_n + \beta} \{ \alpha_n \| x_n - x \|^2 + \beta \| y_n - y \|^2 \} \\ &\leq \max \left(\| x_n - x \|^2, \| y_n - y \|^2 \right), \\ \| y_{n+1} - y \|^2 &\leq \frac{1}{(\mu + \nu_n)(\alpha_n + \beta)} \{ \mu\alpha_n \| x_n - x \|^2 + (\mu\beta + \nu_n(\alpha_n + \beta)) \| y_n - y \|^2 \} \\ &\leq \max \left(\| x_n - x \|^2, \| y_n - y \|^2 \right). \end{aligned}$$

Hence $\max \left(\| x_{n+1} - x \|^2, \| y_{n+1} - y \|^2 \right) \leq \max \left(\| x_n - x \|^2, \| y_n - y \|^2 \right)$. This shows that the sequence (x_n, y_n) is bounded. □

We cannot dispense with assumption $\mathcal{H}47$ now. Note that it is fulfilled if the sequences α_n and ν_n are nonnegative (i.e., $\mathcal{H}44$) and monotonically convergent.

PROPOSITION 4.5. *Assume $\mathcal{H}41$–$\mathcal{H}47$. Then, for any (x, y) inertial Nash equilibrium INE(0,β,μ,0), the sequence (x_n, y_n) generated by (CA^2) satisfies*

(i) $y_n - x_n \to y - x$, $x_{n+1} - x_n \to 0$, $y_{n+1} - y_n \to 0$ *relative to the norm topology in $\mathcal{H}$;*

(ii) $\| x_{n+1} - x \|$ *and* $\| y_{n+1} - y \|$ *have a limit as $n \to +\infty$.*

Proof. Let us adopt some notations:

$$\begin{aligned} A_n &= \alpha_n\mu \| x_n - x \|^2 + \beta(\mu + \nu_n) \| y_n - y \|^2, \\ B_n &= \alpha_n\mu \| x_{n+1} - x_n \|^2 + \beta\nu_n \| y_{n+1} - y_n \|^2 \\ &\quad + \beta\mu \| y_n - x_{n+1} - y + x \|^2 + \beta\mu \| y_{n+1} - x_{n+1} - y + x \|^2 . \end{aligned}$$

So, inequality (c) in Lemma 4.3 reads

$$B_n \leq A_n - \{\alpha_n\mu \| x_{n+1} - x \|^2 + \beta(\mu + \nu_n) \| y_{n+1} - y \|^2\}.$$

Whence we easily derive

$$\begin{aligned} B_n &\leq A_n - A_{n+1} + \mu(\alpha_{n+1} - \alpha_n) \| x_{n+1} - x \|^2 + \beta(\nu_{n+1} - \nu_n) \| y_{n+1} - y \|^2 \\ &\leq A_n - A_{n+1} + [\mu(\alpha_{n+1} - \alpha_n)^+ + \beta(\nu_{n+1} - \nu_n)^+]M, \end{aligned}$$

where M denotes an upper bound for $\| x_{n+1} - x \|^2$ and $\| y_{n+1} - y \|^2$; M is finite in view of proposition 4.4. Define $C_n = [\mu(\alpha_n - \alpha_{n-1})^+ + \beta(\nu_n - \nu_{n-1})^+]M$, and observe

that the series $\sum_{n=1}^{+\infty} C_n$ is convergent owing to assumption $\mathcal{H}47$. The last inequality may be written

$$0 \;\leq\; B_n \;\leq\; \left(A_n - \sum_{k=1}^{n} C_k\right) - \left(A_{n+1} - \sum_{k=1}^{n+1} C_k\right).$$

Hence the sequence $n \mapsto A_n - \sum_{k=1}^{n} C_k$ is decreasing; as it is bounded below by $-\sum_{n=0}^{+\infty} C_n$, it converges. As a consequence A_n converges, too.

Now the inequality above shows that the series $\sum_{k=0}^{+\infty} B_k$ converges, and hence B_n tends to zero as $n \to +\infty$.

Since B_n vanishes as $n \to +\infty$, in particular, $\| y_n - x_{n+1} - y + x \|$ and $\| y_{n+1} - x_{n+1} - y + x \|$ tend to zero. Hence $y_n - x_n - y + x \to 0$, $y_{n+1} - y_n = (y_{n+1} - x_{n+1}) - (y_n - x_{n+1}) \to 0$, $x_{n+1} - x_n = (y_n - x_n) - (y_n - x_{n+1}) \to 0$, which proves point (i).

Further, writing $x_n - x = (y_n - y) - (y_n - x_n - y + x)$ in the expression of A_n yields

$$\begin{aligned} A_n = (\alpha_n\mu + \beta(\mu + \nu_n)) \,\| y_n - y \|^2 \, &- 2\alpha_n\mu\langle y_n - y, y_n - x_n - y + x\rangle \\ &+ \alpha_n\mu \,\| y_n - x_n - y + x \|^2 . \end{aligned}$$

The last two terms vanish as $n \to +\infty$, while $(\alpha_n\mu + \beta(\mu + \nu_n))$ tends to a positive limit. Hence $\| y_n - y \|$ admits a limit as $n \to +\infty$. In view of $x_n - x = (y_n - y) - (y_n - x_n - y + x)$, $\| x_n - x \|$ admits a limit, too. □

Before proceeding to the proof of Theorem 4.1, let us recall a classical argument used to prove the weak convergence of a sequence in a Hilbert space.

LEMMA 4.6 (see Opial [34]). *Let z_n be a sequence in a Hilbert space $\mathcal{H}$ such that there exists a nonvoid set $S \subset \mathcal{H}$ which verifies the following:*

(i) *any weak limit point of z_n belongs to S;*

(ii) $\forall \zeta \in S$, $\lim_{n\to+\infty} \| z_n - \zeta \|$ *exists.*

Then, z_n weakly converges as $n \to +\infty$ to some element of S.

Proof of Theorem 4.1. Recall that the sequences (x_n) and (y_n) are characterized by

$$\begin{cases} 0 \in \partial f(x_{n+1}) + \alpha_n(x_{n+1} - x_n) + \beta(x_{n+1} - y_n), \\ 0 \in \partial g(y_{n+1}) + \mu(y_{n+1} - x_{n+1}) + \nu_n(y_{n+1} - y_n). \end{cases}$$

(a) Let $(\bar{x}, \bar{y})$ be a weak limit point of the bounded sequence (x_n, y_n) (recall Proposition 4.4).

To simplify the notation we suppose for a while that the whole sequence converges. On the one hand $y_n - x_n$ weakly converges to $\bar{y} - \bar{x}$; on the other hand $y_n - x_n$ strongly converges to $y - x$ for any (x, y) inertial Nash equilibrium (Proposition 4.5). Hence $y_n - x_n$ strongly converges to $\bar{y} - \bar{x}$ indeed. Now x_{n+1} satisfies

$$0 \in \partial f(x_{n+1}) + \beta(x_{n+1} - y_{n+1}) + \alpha_n(x_{n+1} - x_n) + \beta(y_{n+1} - y_n).$$

From Proposition 4.5 we know that $x_{n+1} - x_n$ and $y_{n+1} - y_n$ strongly converge to 0; hence $\beta(x_{n+1} - y_{n+1}) + \alpha_n(x_{n+1} - x_n) + \beta(y_{n+1} - y_n)$ strongly converges to $\beta(\bar{x} - \bar{y})$ (α_n is bounded). Owing to the weak-strong closedness of the subgradient operator (see [18]), we have

$$0 \in \partial f(\bar{x}) + \beta(\bar{x} - \bar{y}).$$

Likewise we can prove

$$0 \in \partial g(\bar{y}) + \mu(\bar{y} - \bar{x}).$$

So far we have proved only that any weak limit point $(\overline{x}, \overline{y})$ of the sequence (x_n, y_n) is an inertial Nash equilibrium. But from Proposition 4.5 we know that $\| x_n - x \|$ and $\| y_n - y \|$ have a limit for any (x, y) which is an inertial Nash equilibrium. Opial's lemma (with $\mathcal{H} = \mathcal{X} \times \mathcal{Y}$, $z_n = (x_n, y_n)$ and S denoting the set of inertial Nash equilibrium INE(0,β,μ,0)) then shows that the whole sequence (x_n, y_n) weakly converges to an inertial Nash equilibrium.

(b) The sequence (x_n, y_n) verifies

$$\begin{cases} 0 \in \partial(\mu f)(x_{n+1} + \mu\beta(x_{n+1} - y_{n+1}) + u_n, \\ 0 \in \partial(\beta g)(y_{n+1}) + \beta\mu(y_{n+1} - x_{n+1}) + v_n, \end{cases}$$

where we have put $u_n = \mu\alpha_n(x_{n+1} - x_n) + \mu\beta(y_{n+1} - y_n)$ and $v_n = \beta\nu_n(y_{n+1} - y_n)$. In view of (4.5) this means that $(-u_n, -v_n)$ is a subgradient of $L_{\beta,\mu}$ at point (x_{n+1}, y_{n+1}). Hence, for all $(\xi, \eta) \in \mathcal{H} \times \mathcal{H}$ we have

$$L_{\beta,\mu}(\xi, \eta) \geq L_{\beta,\mu}(x_{n+1}, y_{n+1}) - \langle u_n, \xi - x_{n+1} \rangle - \langle v_n, \eta - y_{n+1} \rangle,$$

but u_n and v_n vanish relative to the norm topology of $\mathcal{H}$ as $n \to +\infty$. Let $(\overline{x}, \overline{y})$ be the weak limit of (x_n, y_n); we then have

$$L_{\beta,\mu}(\xi, \eta) \geq \overline{\lim} L_{\beta,\mu}(x_n, y_n) \geq \underline{\lim} L_{\beta,\mu}(x_n, y_n) \geq L_{\beta,\mu}(\overline{x}, \overline{y}).$$

(c) The strong convergence $y_n - x_n \to \overline{y} - \overline{x}$ has been proved in Proposition 4.5. From point (b) just above we know that $L_{\beta,\mu}(x_n, y_n) = \mu f(x_n) + \beta g(y_n) + \frac{1}{2} \| x_n - y_n \|^2 \to L_{\beta,\mu}(\overline{x}, \overline{y})$. Hence we have

$$\mu f(x_n) + \beta g(y_n) \to \mu f(\overline{x}) + \beta g(\overline{y}),\ n \to +\infty. \tag{4.6}$$

Further,

$$\begin{aligned} \limsup \mu f(x_n) &\leq \overline{\lim}(\mu f(x_n) + \beta g(y_n)) + \overline{\lim}(-\beta g(y_n)) \\ &\leq \mu f(\overline{x}) + \beta g(\overline{y}) - \underline{\lim} \beta g(y_n) \\ &\leq \mu f(\overline{x}), \end{aligned}$$

where the last inequality results from the lower semicontinuity of g. Now the lower semicontinuity of f gives $\overline{\lim} f(x_n) \leq f(\overline{x}) \leq \underline{\lim} f(x_n)$. Finally, with (4.6) we have $g(y_n) \to g(\overline{y})$. □

Remark 1. Condition $\mathcal{H}47$ is but one of various assumptions leading to Theorem 4.1. Indeed, the crux of the matter is inequality (c) of Lemma 4.3.

Let (ρ_n) be a given positive sequence bounded away from zero. Divide each member of inequality (c) in Lemma 4.3 by ρ_n. Proceeding as in the proof of Proposition 4.5, and with the same notation, we can derive

$$\frac{A_{n+1}}{\rho_{n+1}} \leq \frac{A_n}{\rho_n} + \left[\mu \left(\frac{\alpha_{n+1}}{\rho_{n+1}} - \frac{\alpha_n}{\rho_n} \right)^+ + \beta \left(\frac{\mu + \nu_{n+1}}{\rho_{n+1}} - \frac{\mu + \nu_n}{\rho_n} \right)^+ \right] M - \frac{B_n}{\rho_n}.$$

Under the assumption that the series $\sum_{n=0}^{+\infty} \left(\frac{\alpha_{n+1}}{\rho_{n+1}} - \frac{\alpha_n}{\rho_n} \right)^+$ and $\sum_{n=0}^{+\infty} \left(\frac{\mu+\nu_{n+1}}{\rho_{n+1}} - \frac{\mu+\nu_n}{\rho_n} \right)^+$ converge (and also that (α_n) and (ν_n) are bounded), Proposition 4.5 and

Theorem 4.1 can be proved. Relevant choices for ρ_n may be $\rho_n = \alpha_n$, $\rho_n = \nu_n$, $\rho_n = \alpha_n\nu_n$, or $\rho_n = \mu\alpha_n + \beta(\mu + \nu_n)$.

Remark 2. Many interesting questions asking for further developments naturally arise from Theorem 4.1.

Along the same line as in [16], one could try to obtain, by a duality argument, a variational or geometrical characterization of the limits of the sequences generated by our algorithm.

By specializing one of the two functions f or g to be the indicator function of a convex set, one obtains alternating projection-proximal methods for convex programming and variational inequalities (see [38]).

Perspectives. We expect to exploit the simplicity and robustness of such alternating algorithms and apply them to various situations: decomposition and splitting methods in partial differential equations [28, 29], image processing [13, 39], statistics, on-line decision, and so on. Further extensions may concern partial coupling, more than two players, and nonconvex setting (in the line of [5]).

REFERENCES

[1] F. Acker and M.-A. Prestel, *Convergence d'un schéma de minimisation alternée*, Ann. Fac. Sci. Toulouse Math. (5), 2 (1980), pp. 1–9.

[2] S. Adly, H. Attouch, and A. Cabot, *Finite time stabilization of nonlinear oscillators subject to dry friction*, Nonsmooth Mechanics and Analysis, Adv. Mech. Math. 12, Springer, New York, 2006, pp. 289–304.

[3] F. Alvarez, *On the minimizing property of a second order dissipative system in Hilbert spaces*, SIAM J. Control Optim., 38 (2000), pp. 1102–1119.

[4] F. Alvarez and H. Attouch, *An inertial proximal method for maximal monotone operators via discretization of a nonlinear oscillator with damping*, Set-Valued Anal., 9 (2001), pp. 3–11.

[5] H. Attouch and J. Bolte, *On the Convergence of the Proximal Algorithm for Nonsmooth Functions Involving Analytic Features*, Math. Programming B, Nonlinear Convex Optimization and Variational Inequalities, volume in honor of A. Auslender, to appear.

[6] H. Attouch, J. Bolte, and P. Redont, *Optimizing properties of an inertial dynamical system with geometric damping. Link with proximal methods*, Control Cybernet., 31 (2002), pp. 643–657.

[7] H. Attouch, G. Buttazzo, and G. Michaille, *Variational Analysis in Sobolev and BV Spaces. Applications to PDEs and Optimization*, MPS/SIAM Series on Optimization 6, SIAM, Philadelphia, 2006.

[8] H. Attouch, X. Goudou, and P. Redont, *The heavy ball with friction method* I. *The continuous dynamical system: Global exploration of the local minima of a real-valued function by asymptotic analysis of a dissipative dynamical system*, Commun. Contemp. Math., 2 (2000), pp. 1–34.

[9] H. Attouch and A. Soubeyran, *Inertia and Reactivity in Decision Making as Cognitive Variational Inequalities*, Journal of Convex Analysis, 13 n^{o} 2 (2006), pp. 207-224.

[10] H. Attouch and A. Soubeyran, *A Worthwhile to Move Approach of Satisfying with Not Too Much Sacrificing*, J. Math. Psych., submitted.

[11] H. Attouch and M. Teboulle, *Regularized Lotka-Volterra dynamical system as continuous proximal-like method in optimization* , J. Optim. Theory Appl., 121 (2004), pp. 541–570.

[12] J.-P. Aubin and I. Ekeland, *Applied Nonlinear Analysis*, Pure and Applied Mathematics, John Wiley and Sons, New York, 1984.

[13] J.-F. Aujol, G. Aubert, L. Blanc-Féraud, and A. Chambolle, *Image decomposition into a bounded variation component and an oscillating component*, J. Math. Imaging Vision, 22 (2005), pp. 71–88.

[14] H. H. Bauschke, J. M. Borwein, and A. S. Lewis, *The method of cyclic projections for closed convex sets in Hilbert space*, Contemp. Math., 204 (1997), pp. 1–38.

[15] H. H. Bauschke, P. L. Combettes, and D. Noll, *Joint minimization with alternating Bregman proximity operators*, Pac. J. Optim., 2 (2006), pp. 401–424.

[16] H. H. Bauschke, P. L. Combettes, and S. Reich, *The asymptotic behavior of the composition of two resolvents*, Nonlinear Anal., 60 (2005), pp. 283–301.
[17] L. M. Bregman, *The method of successive projections for finding a common point of convex sets*, Sov. Math. Dok., 6 (1984), pp. 699–692.
[18] H. Brézis, *Opérateurs Maximaux Monotones et Semi-groupes de Contractions dans les Espaces de Hilbert et Équations d'Évolution*, North-Holland Mathematics Studies 5, North-Holland, New York, 1973.
[19] R. E. Bruck, *Asymptotic convergence of nonlinear contraction semigroups in Hilbert space*, J. Funct. Anal., 18 (1975), pp. 15–26.
[20] A. Cabot, *Proximal point algorithm controlled by a slowly vanishing term: Applications to hierarchical minimization*, SIAM J. Optim., 15 (2004/05), pp. 555–572.
[21] C. Camerer and G. Loewenstein, *Behavioral Economics: Past, Present, Future*, in Advances in Behavioral Economics, C. Camerer, G. Loewenstein and M. Rabin eds., Princeton University Press, Princeton, NJ, 2003.
[22] P. L. Combettes, *The foundations of set theoretic estimation*, Proceedings of the IEEE, 81 (1993), pp 182–208.
[23] F. Deutsch, *The method of alternating orthogonal projections*, in Approximation Theory, Spline Functions and Applications, S. P. Singh editor, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1992, pp. 105–121.
[24] I. Halperin, *The product of projection operators*, Acta Sci. Math. (Szeged), 23 (1962), pp. 96–99.
[25] J. Hofbauer and S. Sorin, *Best response dynamics for continuous zero-sum games*, Discrete Contin. Dyn. Syst., Ser. B, 6 (2006), pp. 215–224.
[26] H. S. Hundal, *An alternating projection that does not convergence in norm*, Nonlinear Anal., 57 (2004), pp. 35–61.
[27] D. Kahneman, *Maps of Bounded Rationality: Psychology for Behavioral Economics*, American Economic Review, (2003), pp. 1449–1475.
[28] C. Lacour and Y. Maday, *Two different approaches for matching nonconforming grids: The mortar element method and the FETI method*, BIT, 37 (1997), pp. 720–738.
[29] P.-L. Lions, *On the Schwarz alternating method.* III. *A variant for nonoverlapping subdomains*, in Third International Symposium on Domain Decomposition Methods for Partial Differential Equations (Houston, TX, 1989), T. F. Chan, R. Glowinski, J. Périaux, and O. Widlund, eds., SIAM, 1990, pp. 202–231.
[30] P.-L. Lions and B. Mercier, *Splitting algorithms for the sum of two nonlinear operators*, SIAM J. Numer. Anal., 16 (1979), pp 964–979.
[31] E. Matoušková and S. Reich, *The Hundal example revisited*, J. Nonlinear Convex Anal., 4 (2003), pp. 411–427.
[32] J.-J. Moreau, *Proximité et dualité dans un espace hilbertien*, Bull. Soc. Math. France, 93 (1965), pp. 273–299.
[33] J. von Neumann, *Functional Operators.* II. *The Geometry of Orthogonal Spaces*, Annals of Mathematics Studies 22, Princeton University Press, Princeton, NJ, 1950.
[34] Z. Opial, *Weak convergence of the sequence of successive approximations for nonexpansive mappings*, Bull. Amer. Math. Soc., 73 (1967), pp. 591–597.
[35] G. B. Passty, *Ergodic convergence to a zero of the sum of monotone operators in Hilbert space*, J. Math. Anal. Appl., 72 (1979), pp. 383–390.
[36] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.
[37] H. Simon, *A Behavioral Model of Rational Choice*, Quarterly Journal of Economics, 69 (1975), pp. 99–118.
[38] P. Tseng, *Alternating projection-proximal methods for convex programming and variational inequalities*, SIAM J. Optim., 7 (1997), pp 951–965.
[39] L. A. Vese and S. J. Osher, *Modeling textures with total variation minimization and oscillating patterns in image processing*, J. Sci. Comput., 19 (2003), pp. 553–572.

SIAM J. OPTIM.
Vol. 18, No. 3, pp. 1082–1105

TREE APPROXIMATIONS OF DYNAMIC STOCHASTIC PROGRAMS*

RADOSLAVA MIRKOV† AND GEORG CH. PFLUG†

Abstract. We consider a tree-based discretization technique utilizing conditional transportation distance, which is well suited for the approximation of multistage stochastic programming problems, and investigate corresponding convergence properties. We explain the relation between the approximation quality of the probability model and the quality of the solution.

Key words. stochastic programming, multistage, tree approximation, transportation distance, quality of solution

AMS subject classification. 90C15

DOI. 10.1137/060658552

1. Introduction.

Dynamic stochastic optimization models are an up-to-date tool of modern management sciences and can be applied to a wide range of problems, such as financial portfolio optimization, energy contracts, insurance policies, supply chains, etc. While one-period models look for optimal decision values (or decision vectors) based on all available information now, the multiperiod models consider planned future decisions as functions of the information that will be available later. Hence, the natural decision spaces for multiperiod dynamic stochastic models, except for the first stage decisions, are spaces of functions. Consequently, only in some exceptional cases may solutions be found by analytical methods, which include investigating the necessary optimality conditions and solving a variational problem; i.e., only some functional equations have explicit solutions in observed spaces of functions. In the vast majority of cases, it is impossible to find a solution in this way, and we reach out for a numerical solution. However, numerical calculation on digital computers may never represent the underlying infinite-dimensional function spaces. The way out of this dilemma is to approximate the original problem by a simpler, surrogate finite-dimensional problem, which enables the calculation of the numerical solution.

As the optimal decision at each stage is a function of the random components observed so far, the only way to reduce complexity is to reduce the range of the random components. If the random component of the decision model is discrete (i.e., takes a finite and in fact only a few number of values), the variational problem is reduced to a vector optimization problem, which may be solved by well-known vector optimization algorithms. The natural question that arises is how to reconstruct a solution of the basic problem out of the solution of the finite surrogate problem.

Since every finite-valued stochastic process $\tilde{\xi}_1, \ldots, \tilde{\xi}_T$ is representable as a tree, we deal with tree approximations of stochastic processes. There are two contradicting goals to be considered. For the sake of the quality of approximation, the tree should be large and bushy, while for the sake of computational solution effort, it should be small. Thus the choice of the tree size is obtained through a compromise. The basic question in reaching this compromise is the assessment of the quality of the

*Received by the editors April 30, 2006; accepted for publication (in revised form) July 4, 2007; published electronically October 4, 2007.
http://www.siam.org/journals/siopt/18-3/65855.html
†Department of Statistics and Decision Support Systems, University of Vienna, Universitätsstraße 5/3, 1010 Vienna, Austria (radoslava.mirkov@univie.ac.at, georg.pflug@univie.ac.at).

approximation in terms of the tree size. It is the purpose of this paper to shed some light on the relation between the approximation quality of the probability model for the random components and the quality of the solution.

Well-known limiting results are available, which show that by increasing the size of the approximating tree, one eventually gets convergence of the optimal values and the optimizing functions. Results in this direction were proved in [13], [14], etc. Recently, in [9] a stability result has been shown, and an estimate of the approximation error in terms of some distance between the continuous and the discrete models has been given.

Our approach is quite different. We start from the description of the decision model in terms of the system's dynamics functions, the constraint sets, and the objective function. We formulate the whole problem in terms of distributions, as the results are independent from the choice of the probability space. Then we approximate these distributions in an appropriate setting by simpler, discrete distributions and compare the decision functions and the optimal values in both cases.

In our distributional setup, there is no predefined probability space, and therefore there is no room for introducing filtrations other than those which are generated by the random process itself. Thus, we make the following assumption, which is common in stochastic optimization.

Assumption 1. No information other than the values of the scenario process ξ_s, $s \leq t$, is available to the decision maker at time t for $t = 1, \ldots, T$. To put it differently, we assume that the decision x_t is measurable w.r.t. σ-algebra $\mathcal{F}_t$, which is the one generated by $(\xi_1, \ldots, \xi_t)$.

Adopting this approach, there is no need to consider a filtration distance as done in [9].

A further consequence of the in-distribution setting is that we describe the decisions as functions of the random observations and not as functions defined on some probability space. To be more precise, let $\xi^t = (\xi_1, \ldots, \xi_t)$, $t = 1, \ldots, T$, denote the history of the random observations available up to time t. Then the tth decision is a function $\xi^t \mapsto x_t(\xi^t)$ lying in some function space. Noticing that we want to approximate a continuous probability distribution by a discrete one, we note that this function space should respect weak convergence; i.e., if $\xi^{(n)} \to \xi$ weakly, then also $x(\xi^{(n)}) \to x(\xi)$ weakly. Thus we must consider spaces of continuous functions or some subspaces. In this paper, we work with the space of Lipschitz functions. The class of all Lipschitz functions on $\mathbb{R}^n$ determines the weak topology on all probabilities, which have the finite first moment. Moreover, the weakest topology making all integrals of Lipschitz functions continuous is generated by the well-known transportation distance, which is convenient since it can be calculated or at least bounded in many examples.

A result of practical relevance to the decision maker will be shown under some strong regularity assumptions: Suppose that the distance between the original probability model and the approximate one is smaller than ε and that a δ_0-solution of the approximate problem is found. Then one may construct out of it a δ-optimal solution of the original problem. The relation between δ_0, δ, and ε is explicit (see Proposition 4.4). We measure the distance between the original probability model and its discrete approximation by the *conditional transportation distance*, which is finer than the usual *unconditional transportation distance* and accommodates the dynamic character of the problem.

The paper is organized as follows. In section 2 we describe the distance concepts for probability measures and for constraint sets. Section 3 contains the model, and in

section 4 we state the approximation results. Two examples are treated in section 5. In the appendix we have collected some auxiliary results.

2. Preliminaries.

2.1. The probability model. Let $\xi = (\xi_1, \dots, \xi_T)$ be a stochastic process with values in $\Xi \subset \mathbb{R}^{nT}$, and $u = (u_1, \dots, u_T)$ its realization, with $u_t = (u_{t(1)}, \dots, u_{t(n)})$, for each t. Ξ is endowed with the metric

$$d(u_t, v_t) = \sum_{i=1}^{n} |\chi(u_{t(i)}) - \chi(v_{t(i)})|,$$

where χ is a strictly monotonic mapping $\mathbb{R}$ into $\mathbb{R}$. (Ξ, d) is a complete separable metric space such that $\Xi = \Xi_1 \times \cdots \times \Xi_T$, and $\Xi_t \subset \mathbb{R}^n$, for each t. All metrics in all metric spaces appearing in this paper will be denoted by the same symbol d, since there is no danger of confusion. u^t, $t = 1, \dots, T-1$, denotes the history up to time t, i.e., $u^t = (u_1, \dots, u_t)$. Obviously, u^t is an element of the metric space $\Xi^t = \Xi_1 \times \cdots \times \Xi_t \subset \mathbb{R}^{nt}$, which is endowed with the metric

$$d(u^t, v^t) = \sum_{s=1}^{t} d(u_s, v_s).$$

For two Borel measures P, $\tilde{P}$ on a metric space Ξ, we recall that the transportation (Wasserstein) distance d_W between P and $\tilde{P}$ is given by

$$d_W(P, \tilde{P}) = \sup_{f \in Lip_1} \left(\int f(u)\, dP(u) - \int f(u)\, d\tilde{P}(u) \right),$$

where $u \in \Xi$, and Lip_1 is the set of all 1-Lipschitz functions f, i.e.,

$$|f(u) - f(v)| \le d(u, v) \quad \text{for all} \quad u, v \in \Xi.$$

The Wasserstein distance is related to the Monge mass transportation problem (see [20, p. 89]) through the following facts.

- Theorem of Kantorovich and Rubinstein:

  $$\begin{aligned} d_W(P, \tilde{P}) \quad = \quad & \inf\{\mathbb{E}[|Y - \tilde{Y}|], \text{ where the joint distribution } (Y, \tilde{Y}) \\ & \text{is arbitrary, but the marginal distributions are fixed} \\ & \text{such that } Y \sim P;\ \tilde{Y} \sim \tilde{P}\}. \end{aligned}$$

  The infimum here is attained. The optimal joint distribution $(Y, \tilde{Y})$ describes how the mass P should be transported with minimal effort to yield the new mass $\tilde{P}$ (see [20, Theorems 5.3.2 and 6.1.1]).
- For one-dimensional distributions, i.e., distributions on the real line endowed with the Euclidean metric $d(u, v) = |u - v|$, having distribution functions G, resp., $\tilde{G}$, it holds that

  $$d_W(P, \tilde{P}) = \int_{\Omega} |G(u) - \tilde{G}(u)|\, du = \int_0^1 |G^{-1}(u) - \tilde{G}^{-1}(u)|\, du,$$

  where $G^{-1}(u) = \sup\{v : G(v) \le u\}$ (see [25]).

- If χ is a strictly monotonic function mapping $\mathbb{R}$ into $\mathbb{R}$, defining the distance $d(u,v) = |\chi(u) - \chi(v)|$, the pertaining transportation distance is

$$d_W(P, \tilde{P}) = d_W(G \circ \chi^{-1}, \tilde{G} \circ \chi^{-1}) = \int_0^1 |\chi(G^{-1}(u)) - \chi(\tilde{G}^{-1}(u))|\, du.$$

  By an appropriate choice of χ, the convergence of higher moments under d_W-convergence may be ensured (see [16]).

If random variables $Y, \tilde{Y}$ have fixed marginal distributions $P, \tilde{P}$, it makes sense to write

$$d_W(P, \tilde{P}) = d_W(Y, \tilde{Y}) = d_W(G, \tilde{G}), \tag{2.1}$$

as marginal distributions are on par with the specifying random variables, and their distribution functions $G, \tilde{G}$, if they exist (see [27]).

The process ξ generates a Borel probability measure P on Ξ. This measure is characterized by its chain of regular conditional distributions:

$$\begin{aligned} &P(A_1 \times \cdots \times A_T) \\ &= \int_{A_1} \cdots \int_{A_T} P_T(du_T|u^{T-1}) \cdots P_3(du_3|u^2) P_2(du_2|u_1) P_1(du_1) \\ &= \int_{A_1} \cdots \int_{A_T} P_T(du_T|(u_1, \ldots, u_{T-1})) \cdots P_3(du_3|(u_1, u_2)) P_2(du_2|u_1) P_1(du_1). \end{aligned}$$

Here $P_t(A_t|u^{t-1})$ is the conditional probability of $\xi_t \in A_t$ given the past $\xi^{t-1} = u^{t-1}$, $t = 2, \ldots, T$, and P_1 is the probability of $\xi_1 \in A_1$ (not conditional). Since we deal with complete separable metric spaces, the existence of regular conditional probabilities is ensured (see, e.g., [6, Chapter 4, Theorem 1.6]).

The following assumption is imposed on the measure P.

Assumption 2. The conditional probabilities satisfy the Lipschitz condition, i.e.,

$$d_W(P_t(\cdot|u), P_t(\cdot|v)) \le K_t\, d(u,v)$$

for all $u, v \in \Xi^{t-1}$, and some constants K_t, for $t = 2, \ldots, T$.

Remark. Assumption 2 is trivially satisfied if the process is independent. For Markov processes, the condition in Assumption 2 reduces to

$$d_W(P_t(\cdot|u_{t-1}), P_t(\cdot|v_{t-1})) \le K_t\, d(u_{t-1}, v_{t-1}).$$

The ratio

$$\sup_{u,v} \frac{d_W(P(\cdot|u), P(\cdot|v))}{d(u,v)}$$

is called the *ergodic coefficient* of the Markov transition P. If this coefficient is smaller than one, geometric ergodicity holds. Ergodic coefficients were introduced in [4] and extensively applied to stochastic programming problems (see [15]).

Example (Lipschitz condition for multivariate normal distribution). Assume that the process ξ follows the multivariate normal distribution on $\mathbb{R}^T$ equipped with the usual Euclidean metric with mean vector $\mu = \mu^T$ and nonsingular covariance matrix

$\Sigma = \Sigma^T$, where $\mu^t = (\mu_1, \dots, \mu_t)^\top$, and $\Sigma^t = (\sigma_{i,j})$, $i = 1, \dots, t$, $j = 1, \dots, t$, for $t = 1, \dots, T$, are the main submatrices. It holds that

$$\Sigma^t = \begin{bmatrix} \Sigma^{t-1} & \sigma^t \\ (\sigma^t)^\top & \sigma_{t,t} \end{bmatrix}_{t \times t}.$$

Here σ^t, resp., $(\sigma^t)^\top$, denote the tth column, resp., row vector, of the covariance matrix Σ^t of the length $t-1$. According to [11, Theorem 13.1], the conditional distribution of ξ^t given the past realization u^{t-1} is normal with mean $\mu^t + (\sigma^t)^\top (\Sigma^t)^{-1}(u^{t-1} - \mu^{t-1})$ and covariance $\sigma_{t,t} - (\sigma^t)^\top (\Sigma^{t-1})^{-1}\sigma^t$. The transportation distance satisfies

$$d_W(P_t(\cdot|u^{t-1}), P_t(\cdot|v^{t-1})) \le \|(\sigma^t)^\top (\Sigma^t)^{-1}\|_\infty \, d(u^{t-1}, v^{t-1})$$

for $t = 2, \dots, T$, and the constants K_t in Assumption 2 amount to $\|(\sigma^t)^\top (\Sigma^t)^{-1}\|_\infty$.

Let $\tilde{P}$ be some other probability measure that can also be dissected into the chain of conditional probabilities. We introduce the notation

$$\bar{d}_W(P, \tilde{P}) \le (\varepsilon_1, \dots, \varepsilon_T)$$

if

$$\begin{aligned} d_W(P_1, \tilde{P}_1) &\le \varepsilon_1, \\ \sup_{u_1} d_W(P_2(\cdot|u_1), \tilde{P}_2(\cdot|u_1)) &\le \varepsilon_2, \\ \sup_{u^2} d_W(P_3(\cdot|u^2), \tilde{P}_3(\cdot|u^2)) &\le \varepsilon_3, \\ &\vdots \\ \sup_{u^{T-1}} d_W(P_T(\cdot|u^{T-1}), \tilde{P}_T(\cdot|u^{T-1})) &\le \varepsilon_T. \end{aligned}$$

If $\tilde{P}$ is the distribution of a stochastic process $(\tilde{\xi}_1, \dots, \tilde{\xi}_T)$ with finite support, denote by $\tilde{\Xi}^t$ the support of $(\tilde{\xi}_1, \dots, \tilde{\xi}_t)$. The conditional probabilities $\tilde{P}_t(\cdot|u^{t-1})$ are only well defined if $u^{t-1} \in \tilde{\Xi}^{t-1}$. If $u^{t-1} \notin \tilde{\Xi}^{t-1}$, we set $d_W(P_t(\cdot|u^{t-1}), \tilde{P}_t(\cdot|u^{t-1})) = 0$, which is the same as saying that the conditional probabilities of $\tilde{P}$ are set equal to those of P, for those values, which will not be taken by $\tilde{P}$ with positive probability.

Example (convergence of tree processes). We consider the tree process as in [9] shown in Figure 2.1.

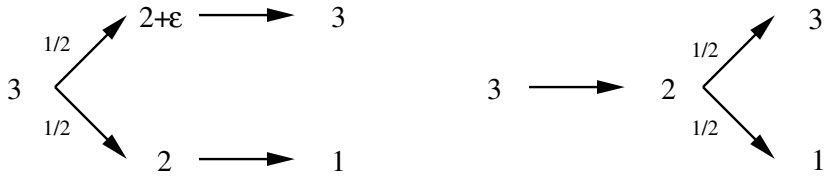


FIG. 2.1. *Tree processes and their conditional distributions.*

The processes describe the development of prices at time $t = 1, 2, 3$ used to determine an optimal purchase strategy over time under cost uncertainty by the means of multistage stochastic programming methods. The stochastic price processes as in Figure 2.1 (left) yield the probability distribution $P = (P_1, P_2, P_3)$ for $\varepsilon \in [0, 1)$, and

for $\varepsilon = 0$ we obtain the approximation $\tilde{P}$ of P (right). The filtrations of σ-fields generated by ξ and $\tilde{\xi}$ do not coincide.

Although the processes on the left obviously converge in distribution to the process on the right, as ε tends to zero, these processes are far apart in conditional transportation distance. First, the processes on the left do not satisfy the condition of Assumption 2 uniformly in ε, as

$$d_W(P_3(\cdot|(3,2)), P_3(\cdot|(3,2+\varepsilon))) = 2 = \frac{2}{\varepsilon} d(2, 2+\varepsilon).$$

Moreover, no convergence in the conditional transportation distance to the process on the right holds as ε tends to zero. A closer look at $d_W(P_t(\cdot|u^{t-1}), \tilde{P}_t(\cdot|u^{t-1}))$, $t = 1, 2, 3$, shows that

$$d_W(P_1, \tilde{P}_1) = 0,$$
$$d_W(P_2(\cdot|3), \tilde{P}_2(\cdot|3)) = \frac{\varepsilon}{2} \to 0,$$

but

$$d_W(P_3(\cdot|(3,2)), \tilde{P}_3(\cdot|(3,2))) = \frac{1}{2} d(3,1) = 1,$$
$$d_W(P_3(\cdot|(3,2+\varepsilon)), \tilde{P}_3(\cdot|(3,2+\varepsilon))) = 0,$$

and $\sup_{u^2} d_W(P_3(\cdot|u^2), \tilde{P}_3(\cdot|(u^2) = 1$; i.e., it does not tend to 0.

Associated with the process of observation ξ_t is the the process of decisions x_t, where x_t are continuous mappings from Ξ^t into $\mathbb{R}^{m_t}$. On $\mathbb{R}^m$, $m = \sum_{t=1}^T m_t$, we work with an appropriate distance d. We will assume that the optimal decisions are Lipschitz continuous. The next example shows why it is hopeless to get rid of assumptions on the smoothness of the solutions.

Example (mean absolute deviation regression). Let ξ_1 be a uniform $[0,1]$ variable and let ξ_2, conditional on ξ_1, have a normal distribution with mean $\xi_1/2$ and variance 1. Denote the distribution of (ξ_1, ξ_2) by P.

We want to solve

$$\min_{x(\xi_1)} \mathbb{E}_P(|x(\xi_1) - \xi_2|), \tag{2.2}$$

where $x(\xi_1)$ is a measurable function of ξ_1. Obviously, the solution is

$$x(u) = u/2 \tag{2.3}$$

since the normal distribution is symmetric around zero. However, consider a sequence of discrete measures $\tilde{P}^{(n)}$, converging weakly to P, and assume that these measures sit on $(\xi_{1,i}^{(n)}, \xi_{2,i}^{(n)})$ with equal probability $1/n$, such that all $\xi_{1,i}^{(n)}$ are distinct. Then the solutions of

$$\min_{x(\xi_1)} \frac{1}{n} \sum_{i=1}^n (|x(\xi_{1,i}^{(n)}) - \xi_{2,i}^{(n)}|) \tag{2.4}$$

would not converge to (2.3) in any sense.

Smoothing techniques (sieve techniques) are used in nonparametric regression for ensuring consistency. This technique would for fixed n search a solution $x(\cdot)$ in

a smooth function space and gradually but slowly increase the function space as n increases. The convergence rate depends on the degree of smoothness of the solution. To put this in terms of optimization, the approximate problem (2.4) is solved in nonparametric statistics under an additional constraint of smoothness, which is not present in the original problem (2.2). Sieve techniques were introduced by [2] and are standard in nonparametric curve estimation (see, for instance, [7]). Only in rare cases is the additional smoothness condition superfluous because of strong shape conditions [21], [3].

While in statistical applications one has to take the data as observed, one may choose the approximating model in stochastic optimization. In other words, and this is the approach we adopt here, one may choose the approximating model such that also the conditional probabilities of the approximations are close to the original ones. Let us see how this would solve the above nonparametric regression problem. Suppose we choose the points $i/(n+1)$, $i = 1, \dots, n$, each with probability $1/n$, as a good approximation of the uniform distribution in the first component. Then choose conditional on $\xi_1 = i/(n+1)$ a discrete distribution which would come close to the original normal distribution by a transportation distance not larger than ε. By doing so, also the median would not differ more than ε, and by a simple linear interpolation between the points sitting at $i/(n+1)$ and at $(i+1)/(n+1)$ we would have found the true regression line within a sup-norm distance of ε, too.

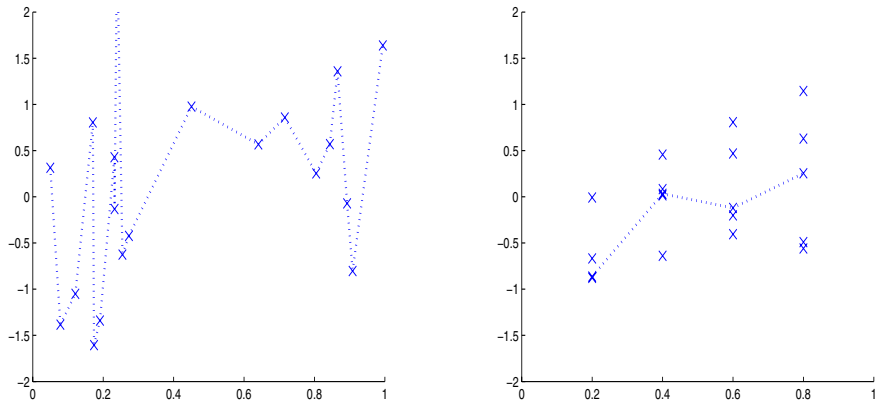


FIG. 2.2. *Discrete approximations of the distribution of* (ξ_1, ξ_2). *The distribution on the left is close in transportation distance, while the distribution on the right is in addition close in conditional transportation distance.*

For an illustration, see Figure 2.2. The left-hand side shows a sample of the two-dimensional distribution which approximates the uniform distribution on the square based on the transportation distance. On the right-hand side, we have chosen the x-coordinates as $i/5$, $i = 1, \dots, 4$, and sampled only the conditional distributions; i.e., we have a tree-structured distribution, which approximates in addition some conditional distributions. In both cases we have shown the (linearly interpolated) solution of problem (2.2) as a dotted line.

2.2. The projection distance. Let $\mathbb{B}(r) = \{x \in \mathbb{R}^m : d(0, x) \le r\}$ denote the ball with diameter $2r$ in $\mathbb{R}^m$. The projection distance $d_{P,r}$ between two closed, convex

sets $A, B \subseteq \mathbb{R}^m$ is defined as

$$d_{P,r}(A,B) = \sup_{x \in \mathbb{B}(r)} d(proj_A(x), proj_B(x)),$$

where $proj_A(x)$ denotes the convex projection of the point x onto A. We allow r to take the value ∞ and set

$$d_P(A,B) = d_{P,\infty}(A,B) = \sup_{x \in \mathbb{R}^m} d(proj_A(x), proj_B(x)).$$

The projection distance is larger than the usual Hausdorff distance, which is

$$d_H(A,B) = \max(\sup_{x \in A} d(x,B), \sup_{x \in B} d(x,A)),$$

since

$$d_H(A,B) = \sup_{x \in A \cup B} d(proj_A(x), proj_B(x)) \leq d_{P,\infty}(A,B).$$

Example (projection and Hausdorff distance). Let us show that there is no converse Lipschitz relation between d_H and $d_{P,\infty}$. Consider in $\mathbb{R}^2$ as set A the line segment connecting $[-1, \varepsilon]$ and $[1, -\varepsilon]$ and as B the line segment connecting $[-1, -\varepsilon]$ and $[1, \varepsilon]$. Here the Euclidean distance is used. A and B are closed convex sets with Hausdorff distance $d_H(A,B)$. However, choosing the point $x = (0, \varepsilon + 1/\varepsilon)$, one gets that the Hausdorff distance is smaller by an order than the projection distance, since $d_P(A,B)$ is the hypotenuse of a triangle whose one leg is $d_H(A,B)$. More precisely, $d_P(A,B) = 2$, and for $\varepsilon \in (0,1]$, $d_H(A,B) \leq 2\varepsilon$. In Figure 2.3 the corresponding distances for $\varepsilon = 1/2$ are represented by dotted lines.

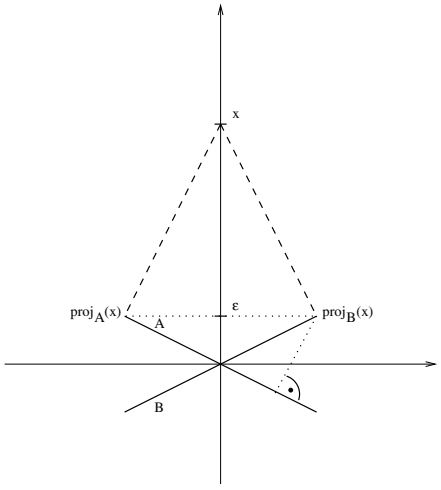


FIG. 2.3. *The Hausdorff and the projection distance.*

Example (Lipschitzian property of projection distance). Let A_1 and A_2 be the hyperplanes

$$A_i = \{x : s_i^\top x = w_i\} \subset \mathbb{R}^m, \qquad i = 1, 2.$$

Then

$$proj_{A_i}(x) = x + \frac{(w_i - s_i^\top x)s_i}{\|s_i\|^2}$$

and

$$\|proj_{A_1}(x) - proj_{A_2}(x)\|$$

$$\leq \left(|w_1| + \|s_1\|\|x\| + \frac{|w_2|}{\|s_2\|}(\|s_1\| + \|s_2\|)\right)\frac{\|s_1 - s_2\|}{\|s_1\|^2} + |w_1 - w_2|\frac{\|s_2\|}{\|s_1\|^2} + \|x\|\frac{\|s_1\| + \|s_2\|}{\|s_1\|}.$$

Suppose that s as well as w depend in a Lipschitz way on a parameter u. Then the set-valued mapping

$$u \mapsto \{x \in \mathbb{B}(r) : s(u)^\top x = w(u)\}$$

for $r < \infty$ is Lipschitz w.r.t. the projection distance $d_{P,r}$, if $\|s(u)\|$ and $|w(u)|$ are bounded, and $\|s(u)\|$ is bounded away from zero.

We utilize the projection distance in assumptions for the behavior of the constraint set.

Remark. Note that one could still use the weaker concept of the Hausdorff distance d_H, which would put the results in a more classical setting, and allow the use of already existing results (see, e.g., [23]). In that case there is no Lipschitz continuity, and at most (sub-)Hölder continuity can be obtained. If $\mathcal{X}(z)$ denotes the constraint set (for the definition of and assumptions about constraint sets see section 3), we say that d_r is sub-Hölder continuous with modulus α if, for each r, there exists a constant M_r such that

$$d_r(\mathcal{X}(z), \mathcal{X}(\bar{z})) \leq M_r[d(z, \bar{z})]^\alpha,$$

where d_r denotes the r-distance between closed, convex sets A and B, i.e.,

$$d_r(A, B) = \sup_{x \in \mathbb{B}(r)} |d(x, A) - d(x, B)|.$$

In our case we obtain sub-Hölder behavior of constraint sets with $\alpha = 1/2$ and $M_r = 2\sqrt{r}$ (see Lemma A.1). Still, Hölder property is weaker than Lipschitz and does not propagate well in multiperiod situations. For this reason we stick to the projection distance.

3. Dynamic decision models. We represent the multistage dynamic decision model as a state-space model. We assume that there is a state vector ζ_t, which describes the situation of the decision maker at time t immediately before he must make the decision x_t, for each $t = 1, \ldots, T$, and its realization is denoted by z_t. The initial state $\zeta_0 = z_0$, which precedes the deterministic decision at time 0, is known and is given by ξ_0. To assume the existence of such a state vector is no restriction at all, since we may always take the whole observed past and the decisions already made as the state:

$$\zeta_t = (x^{t-1}, \xi^t), \qquad t = 1, \ldots, T.$$

However, the vector of required necessary information for future decisions is often much shorter.

The state variable process $\zeta = (\zeta_1, \dots, \zeta_T)$, with realizations $z = (z_1, \dots, z_T)$, is a controlled stochastic process, which takes values in a metric state space $Z = Z_1 \times \cdots \times Z_T$. The control variables are the decisions x_t, $t = 1, \dots, T$. The state ζ_t at time t depends on the previous state ζ_{t-1}, the decision x_{t-1} following it, and the last observed scenario history ξ^t with realization u^t. A transition function g_t describes the state dynamics:

$$\zeta_t = g_t(\zeta_{t-1}, x_{t-1}, \xi^t), \qquad t = 1, \dots, T. \tag{3.1}$$

At the terminal stage T, no decisions are made; only the outcome $\zeta_T = z_T$ is observed.

Note that ζ_t, as a function of the random variable ξ^t, is a random variable with realization z_t, which is a function of u^t, the realization of ξ^t, for each $t = 1, \dots, T$.

The decision x of the multistage stochastic problem is a vector of continuous functions $x = (x_0, \dots, x_{T-1})$, where x_t, $t = 1, \dots, T-1$, maps Ξ^t to $\mathbb{R}^{m_t}$. We require that the feasible decision x_t at time t satisfies a constraint of the form

$$x_t \in \mathcal{X}_t(\zeta_t), \qquad t = 1, \dots, T,$$

where $\mathcal{X}_t$ are closed convex multifunctions with closed convex values. Let us now define a Ξ-feasible decision.

DEFINITION 3.1. *We say that x is a Ξ-feasible decision if the following are fulfilled:*

1. $x_0 \in \mathcal{X}_0(z_0)$;
2. $u \mapsto x_t(u)$ *is a continuous function* $\Xi^t \to \mathbb{R}^{m_t}$, $t = 1, \dots, T-1$;
3. *if, for* z_0 *given and* $(u_1, \dots, u_T) \in \Xi$, z_t *is recursively defined by*

$$z_t(u^t) = g_t(z_{t-1}(u^{t-1}), x_{t-1}(u^{t-1}), u^t), \qquad t = 1, \dots, T,$$

then

$$x_t(u^t) \in \mathcal{X}_t(z_t(u^t)), \qquad t = 1, \dots, T-1.$$

$\mathcal{X}$ denotes the set of all Ξ-feasible decisions $x = (x_0, x_1(\xi^1), \dots, x_{T-1}(\xi^{T-1}))$.

The objective to be minimized is

$$F(x, P) := \sum_{t=1}^{T} \mathbb{F}_t[\zeta_t], \tag{3.2}$$

where $\mathbb{F}_t$ are version-independent probability functionals, i.e., mappings from a space of random variables on Z_t to the real line, where the function values depend only on the distribution and not on the concrete version of the random variable. Examples for version-independent probability functionals are the expectation, the moments, the mean absolute deviation, and all typical risk functionals used in finance.

Finally, we obtain the multistage problem in the state-dynamics representation:

$$\begin{array}{ll} \text{minimize in } x & F(x, P) \\ \text{subject to} & x \in \mathcal{X}, \end{array} \tag{3.3}$$

i.e.,

$$\begin{array}{ll} \text{minimize in } x & \displaystyle\sum_{t=1}^{T} \mathbb{F}[\zeta_t] \\ \text{subject to} & x \in \mathcal{X}; \\ & \zeta_t \text{ is obtained through the recursion (3.1).} \end{array}$$

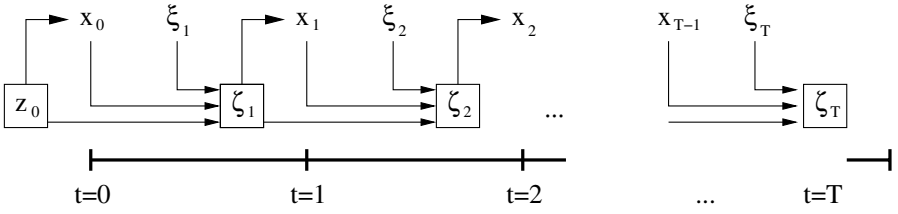


FIG. 3.1. *Multistage state-dynamics decision process.*

Its graphical representation is given in Figure 3.1.

Since the decisions x_t of the multistage problem are functions of the states ζ_t, the optimal set of decisions is expressed as a set of functions $x_t = x_t(\zeta_t)$. The problem (3.3) is a variational problem, and explicit solution methods for it exist only in exceptional cases.

Remark. The existence of continuous solutions (or better continuous selections from the arg min-sets) in our setting requires further consideration. We just mention here that such a type of result is contained in [22], where it is shown that under continuity and convexity assumptions, the existence of continuous measurable nonanticipative solutions is guaranteed if the probability measure is laminary. The property of laminarity is close to our Assumption 2 but is not implied by it. To be more precise, assume that $S_t(u^t)$ is the support of the conditional distribution of $(\xi_{t+1}, \dots, \xi_T)$ given $\xi^t = u^t$. It is not difficult to see that Assumption 2 implies that $u^t \mapsto S_t(u^t)$ is a closed-valued, lower semicontinuous multifunction. If $u^t \mapsto S_t(u^t)$ is, in addition, continuous (in the Painlevé–Kuratowski sense) for $t = 1, \dots, T$, then the distribution of ξ is laminary.

We now state a set of smoothness and dependence assumptions for the model.

Assumption 3. Let for each $t = 1, \dots, T$, and some real constants L_t, M_t, N_t, the following hold.

- The functions g_t satisfy

$$d\left(g_t(z,x,u), g_t(\bar{z},\bar{x},\bar{u})\right) \le L_t\left(d(z,\bar{z}) + d(x,\bar{x}) + d(u,\bar{u})\right)$$

  for $z, \bar{z}$ with values in Z_{t-1}, $x \in \mathcal{X}_{t-1}(z)$, $\bar{x} \in \mathcal{X}_{t-1}(\bar{z})$, $u, \bar{u} \in \Xi^t$.
- The constraints are described by closed convex sets $\mathcal{X}_t(z)$, which depend in a Lipschitz way on z, such that

$$d_{P,r}(\mathcal{X}_t(z), \mathcal{X}_t(\bar{z})) \le M_t d(z,\bar{z}),$$

  for $z, \bar{z}$ with values in Z_t, where $d_{P,r}$ denotes the projection distance between closed convex sets. Here r is finite if we know that the solutions lie in $\mathbb{B}(r)$; otherwise we set $r = \infty$.
- The version-independent probability functionals $\mathbb{F}_t$ satisfy

$$|\mathbb{F}_t(\zeta) - \mathbb{F}_t(\bar{\zeta})| \le N_t d_W(\zeta,\bar{\zeta}),$$

  where d_W is the Wasserstein distance, and $\zeta, \bar{\zeta}$ with distributions on Z_t.

In view of (2.1), we write $d_W(\zeta,\bar{\zeta})$ instead of $d_W(P,\bar{P})$ for some adequate probability measure $\bar{P}$.

Remark. The second Lipschitz condition of Assumption 3 assures that $\mathcal{X}_t(z)$ is a nonempty set, for all z with values in Z_t, for $t = 1, \dots, T$; i.e., no induced constraints are allowed.

Assumption 4. The components $\xi_t = (\xi_{t(1)}, \dots, \xi_{t(n)})$ of the random process ξ are conditionally independent given the past ξ^{t-1}.

This assumption is not as strong as it sounds. In the dynamics

$$\zeta_t = g_t(\zeta_{t-1}, x_{t-1}, \xi^{t-1}, \xi_{t(1)}, \dots, \xi_{t(n)}),$$

we may assume that the risk factors $(\xi_{t(1)}, \dots, \xi_{t(n)})$ are conditionally independent given the past ξ^{t-1}; otherwise we transform the originally dependent components into conditionally independent ones and reformulate the transition function g_t accordingly.

4. Approximations. Instead of the original problem (3.3) we consider a tree process $(\tilde{\xi}_1, \dots, \tilde{\xi}_T)$ with distribution $\tilde{P}$ and support $\tilde{\Xi}$. The decision based on the approximation is $\tilde{\Xi}$-feasible. We assume that $\tilde{\Xi} \subset \Xi$.

DEFINITION 4.1. *We say that x is a $\tilde{\Xi}$-feasible decision if the following are fulfilled:*

1. $x_0 \in \tilde{\mathcal{X}}_0(\tilde{z}_0)$;
2. $\tilde{u} \mapsto x_t(\tilde{u})$ *is a continuous function* $\tilde{\Xi}^t \to \mathbb{R}^{m_t}$, $t = 1, \dots, T-1$;
3. *if, for* $\tilde{z}_0$ *given and* $(\tilde{u}_1, \dots, \tilde{u}_T) \in \tilde{\Xi}$, $\tilde{z}_t$ *is recursively defined by*

$$\tilde{z}_t(\tilde{u}^t) = g_t(\tilde{z}_{t-1}(\tilde{u}^{t-1}), x_{t-1}(\tilde{u}^{t-1}), \tilde{u}^t), \qquad t = 1, \dots, T,$$

then

$$x_t(\tilde{u}^t) \in \tilde{\mathcal{X}}_t(\tilde{z}_t(\tilde{u}^t)), \qquad t = 1, \dots, T-1.$$

$\tilde{\mathcal{X}}$ *denotes the set of all* $\tilde{\Xi}$*-feasible decisions* $x = (x_0, x_1(\tilde{\xi}^1), \dots, x_{T-1}(\tilde{\xi}^{T-1}))$.

This yields the approximate problem

$$\begin{array}{ll} \text{minimize in } x & F(x, \tilde{P}) \\ \text{subject to} & x \in \tilde{\mathcal{X}}, \end{array} \tag{4.1}$$

i.e.,

$$\begin{array}{ll} \text{minimize in } x & \displaystyle\sum_{t=1}^{T} \mathbb{F}[\tilde{\zeta}_t] \\ \text{subject to} & x \in \tilde{\mathcal{X}}; \\ & \tilde{\zeta}_t \text{ is obtained through the recursion (3.1).} \end{array}$$

Remark. We do not propose the use of any random sampling technique to construct the tree but rather to control the transportation distance between the conditional distributions of the original problem and of the approximating tree. In particular, the first approximation is obtained by making $d_W(P_1, \tilde{P}_1)$ small. This amounts to finding a good solution of a facility location problem (see [10]). Suppose that n_1 points $u_{1(1)}, \dots, u_{1(n_1)}$ on the first stage have been selected. Then we choose the conditional probabilities $\tilde{P}(\cdot|u_{i(j)})$ such that they are close to $P(\cdot|u_{i(j)})$, $j = 1, \dots, n_1$, again in transportation distance. This procedure is then repeated through all stages.

The next two propositions tell us how to arrive from a solution of the original problem to a solution of the approximate problem, and vice versa. We assume that both problems have Lipschitz solutions. In section 5, we give examples, which show that the Lipschitz property of the solution is quite natural in problems from finance and supply chain management. In the inventory control problem the solution is Lipschitz continuous if the dependence between the under-/overshooting quantity and costs is linear, which is typically fulfilled.

PROPOSITION 4.2 (restriction). *Suppose that Assumption* 3 *holds, and*

$$\bar{d}_W(P, \tilde{P}) \le (\varepsilon_1, \dots, \varepsilon_T),$$

where P is supported by Ξ, and $\tilde{P}$ is supported by the finite set $\tilde{\Xi}$. Then every Ξ-feasible decision x, which is Q-Lipschitz, is also $\tilde{\Xi}$-feasible, and we have that

$$|F(x, P) - F(x, \tilde{P})| \le \delta_1, \tag{4.2}$$

with

$$\delta_1 = \sum_{s=1}^{T} \bar{\varepsilon}_s \sum_{t=s}^{T} N_t D_{t,s},$$

where

$$\bar{\varepsilon}_s = \sum_{i=1}^{s} \varepsilon_i \prod_{j=i+1}^{s} K_j, \tag{4.3}$$

and, for $Q = (Q_0, \dots, Q_{T-1})$, the constants $D_{s,t}$ fulfill the recursion

$$\begin{aligned} D_{t,t} &= L_t, \\ D_{t,s} &= L_t D_{t-1,s} + Q_{t-1}, \qquad s = 1, \dots, t-1. \end{aligned} \tag{4.4}$$

Proof. It is evident that every Ξ-feasible decision is automatically $\tilde{\Xi}$-feasible.

Under Assumption 2, one may construct for every t stochastic processes $\xi_t(u)$, $u \in \Xi^{t-1}$, and $\tilde{\xi}_t(u)$, $u \in \tilde{\Xi}^{t-1}$, sitting for every t on independent (product) probability spaces such that

(i) $\xi_t(u)$ has distribution $P_t(\cdot|u)$;

(ii) for all $u, v \in \Xi^{t-1}$,

$$\mathbb{E}[d(\xi_t(u), \xi_t(v))] = d_W(P_t(\cdot|u), P_t(\cdot|v)) \le K_t\, d(u, v);$$

(iii) if $u \in \tilde{\Xi}^{t-1}$, then

$$\mathbb{E}[d(\xi_t(u), \tilde{\xi}_t(u))] = d_W(P_t(\cdot|u), \tilde{P}_t(\cdot|u)) \le \varepsilon_t.$$

The construction of the processes is based on the following. Since the components $\xi_t = (\xi_{t,1}, \dots, \xi_{t,n})$ of the process ξ are conditionally independent (Assumption 4), their conditional distribution functions at $(y_1, \dots, y_n)$ given $u^{t-1} = u$ can be written as $G_{t,u,1}(y_1), \dots, G_{t,u,n}(y_n)$. Then $\xi_t(u)$ is defined as

$$\xi_t(u) = (G^{-1}_{t,u,1}(U_1), \dots, G^{-1}_{t,u,n}(U_n)),$$

where $U_1, \dots, U_n$ are independent and identically distributed Uniform$[0, 1]$. By this construction, (i) and (ii) are fulfilled. As to (iii), recall the theorem of Kantorovich and Rubinstein, and notice that the infimum is attained. This joint "minimal" distribution can be glued to the distribution $P_t(\cdot|u)$ to entail (iii) (see [26, Lemma 7.6]).

The processes ξ_t and $\tilde{\xi}_t$ then appear as compositions $\xi_t(\xi_{t-1}(\cdots(\xi_1)\cdots))$, resp., $\tilde{\xi}_t(\tilde{\xi}_{t-1}(\cdots(\tilde{\xi}_1)\cdots))$. We show that $\mathbb{E}[d(\xi_t, \tilde{\xi}_t)] \le \bar{\varepsilon}_t$. Obviously, $\mathbb{E}[d(\xi_1, \tilde{\xi}_1)] \le \varepsilon_1$. Suppose that it is already shown that $\mathbb{E}[d(\xi_{t-1}, \tilde{\xi}_{t-1})] \le \bar{\varepsilon}_{t-1}$. Then

$$\begin{aligned} \bar{\varepsilon}_t &= \mathbb{E}[d(\xi_t(\xi_{t-1}), \tilde{\xi}_t(\tilde{\xi}_{t-1}))] \\ &\le \mathbb{E}[d(\xi_t(\xi_{t-1}), \xi_t(\tilde{\xi}_{t-1}))] + \mathbb{E}[d(\xi_t(\tilde{\xi}_{t-1}), \tilde{\xi}_t(\tilde{\xi}_{t-1}))] \\ &\le K_t \bar{\varepsilon}_{t-1} + \varepsilon_t, \end{aligned}$$

leading to (4.3).

Now we argue pointwise for a specific ω in the standard probability space $[0,1]^{\mathbb{N}}$ with Lebesgue measure, which we have constructed. All of the following calculations are done for this specific ω. We use the fact that a pointwise argument implies a distributional result.

Setting $d(\xi_t, \tilde{\xi}_t) = d_t$ and $d(\xi^t, \tilde{\xi}^t) = d^t = \sum_{s=1}^t d_s$,

$$d(\zeta_1, \tilde{\zeta}_1) = d(g_1(z_0, x_0, \xi_1), g_1(z_0, x_0, \tilde{\xi}_1)) \le L_1 d^1$$

implies $D_{1,1} = L_1$. Suppose that

$$d(\zeta_{t-1}, \tilde{\zeta}_{t-1}) \le \sum_{s=1}^{t-1} D_{t-1,s} d_s.$$

Then

$$\begin{aligned} d(\zeta_t, \tilde{\zeta}_t) &= d(g_t(\zeta_{t-1}, x_{t-1}(\xi^{t-1}), \xi^t), g_t(\tilde{\zeta}_{t-1}, x_{t-1}(\tilde{\xi}^{t-1}), \tilde{\xi}^t)) \\ &\le L_t \left(\sum_{s=1}^{t-1} D_{t-1,s} d_s + Q_{t-1} \sum_{s=1}^{t-1} d_s + \sum_{s=1}^{t} d_s \right) \\ &= \sum_{s=1}^{t} D_{t,s} d_s, \end{aligned}$$

with

$$\begin{aligned} &D_{t,t} = L_t, \\ &D_{t,s} = L_t D_{t-1,s} + Q_{t-1}, \qquad s = 1, \dots, t-1. \end{aligned}$$

Therefore, taking the expectations,

$$\begin{aligned} \left| \sum_{t=1}^{T} (\mathbb{F}_t[\zeta_t] - \mathbb{F}[\tilde{\zeta}_t]) \right| &\le \sum_{t=1}^{T} N_t d_W(\zeta_t, \tilde{\zeta}_t) \\ &\le \sum_{t=1}^{T} N_t \mathbb{E}[d(\zeta_t, \tilde{\zeta}_t)] \\ &\le \sum_{t=1}^{T} N_t \sum_{s=1}^{t} \bar{\varepsilon}_s D_{t,s} \\ &= \sum_{s=1}^{T} \bar{\varepsilon}_s \sum_{t=1}^{T} N_t D_{t,s}, \end{aligned}$$

which yields (4.2). □

PROPOSITION 4.3 (extension). *Suppose that Assumption* 3 *holds, and*

$$\bar{d}_W(P, \tilde{P}) \le (\varepsilon_1, \dots, \varepsilon_T),$$

where P is supported by Ξ, and $\tilde{P}$ is supported by the finite set $\tilde{\Xi}$. Then for every $\tilde{\Xi}$-feasible decision $\tilde{x}$ of (3.3)*, which is Q-Lipschitz, there is a Ξ-feasible decision x, which is Q^e-Lipschitz, called the extension of $\tilde{x}$, such that*

$$|F(x, P) - F(\tilde{x}, \tilde{P})| \le \delta_2, \tag{4.5}$$

where

$$\delta_2 = \sum_{s=1}^{T} \bar{\varepsilon}_s \sum_{t=s}^{T} N_t D^e_{t,s}.$$

The $\bar{\varepsilon}$'s are given by (4.3) *and for $Q^e = (Q^e_0, \dots, Q^e_{T-1})$, the constants Q^e_t, and $D^e_{s,t}$ fulfill the recursions*

$$\begin{aligned} Q^e_t &= Q_t + M_t \sum_{s=1}^{t-1} D^e_{t-1,s}, \\ D^e_{t,t} &= L_t, \\ D^e_{t,s} &= L_t D^e_{t-1,s} + Q^e_{t-1}, \qquad s = 1, \dots, t-1. \end{aligned}$$

Proof. Again, we construct the processes ξ and $\tilde{\xi}$ on the standard probability space $[0,1]^{\mathbb{N}}$. As in Proposition 4.2, the argumentation is pointwise for a specific ω in this probability space.

Set $d(\xi_t, \tilde{\xi}_t) = d_t$, and $d(\xi^t, \tilde{\xi}^t) = d^t = \sum_{s=1}^{t} d_s$.

Let $\tilde{x}$ be a Q-Lipschitz, $\tilde{\Xi}$-feasible family of decisions. By the extension theorem (Theorem A.2), we may extend these functions to a family of functions x^e defined on the whole Ξ with the same Lipschitz constant.

It may happen that the x^e functions are not feasible. We have to make them feasible in a recursive way, starting with x^e_1, then x^e_2, and so on.

Let $\zeta_1(\xi_1) = g_1(\tilde{z}_0, \tilde{x}_0, \xi_1)$. Then

$$d(\zeta_1(\xi_1), \tilde{\zeta}_1(\tilde{\xi}_1)) \le L_1 d^1.$$

Now let

$$x_1(\xi_1) := proj_{\mathcal{X}(z_1)}(x^e_1(\xi_1)).$$

Since $\mathcal{X}(z_1(u))$ is $(M_1 L_1)$-Lipschitz and x^e_1 is Q_1-Lipschitz, we have by Lemma A.3 that x_1 is $(Q_1 + M_1 L_1)$-Lipschitz and that

$$d(\zeta_2, \tilde{\zeta}_2) \le L_2(L_1 d_1 + (Q_1 + M_1 L_1) d_1 + d_2).$$

This argument gets recursively iterated as in the proof of Proposition 4.2, leading to the indicated sequences of constants and finally to (4.5). □

Remark. There are various variants of Proposition 4.3, for instance, if norms are replaced by equivalent ones or if the process ξ is Markovian. For particular models, much finer and better estimates may be found.

Recall the notion of δ-optimal solutions. In the given setting, they belong to

$$\begin{aligned} \delta\text{-}\arg\min_x F(x,P) &= \{\bar{x} \in \mathcal{X} \mid F(\bar{x},P) \le \inf_x F(x,P) + \delta\}, \quad \text{resp.}, \\ \tilde{\delta}\text{-}\arg\min_x F(x,\tilde{P}) &= \{\bar{x} \in \mathcal{X} \mid F(\bar{x},\tilde{P}) \le \inf_x F(x,\tilde{P}) + \tilde{\delta}\}. \end{aligned}$$

The concept of approximately optimal solutions offers us more flexible framework, as the δ-$\arg\min$-mappings satisfy the Lipschitz continuity under some additional conditions; i.e., F has to be proper, lower semicontinuous, and convex (cf. [23, Chapter 7J]).

Based on the statements above, we have the following result.

PROPOSITION 4.4. *Suppose that Assumption* 3 *holds, and*

$$\bar{d}_W(P, \tilde{P}) \le (\varepsilon_1, \dots, \varepsilon_T),$$

where P is supported by Ξ, and $\tilde{P}$ is supported by the finite set $\tilde{\Xi}$, and the approximation has been chosen so that the values of $\delta_1 = \delta_1(\varepsilon)$ and $\delta_2 = \delta_2(\varepsilon)$ are given, for $\varepsilon = (\varepsilon_1, \dots, \varepsilon_T)$. If $\tilde{x}^*$ is a δ_0-optimal solution of the approximate problem (4.1)*, then its extension x^* is a $(\delta_0 + \delta_1 + \delta_2)$-optimal solution of the basic problem* (3.3).

Proof. Let y^* be a solution of the basic problem (3.3). Since $\tilde{x}^*$ is a δ_0-solution of the approximate problem (4.1),

$$F(\tilde{x}^*, \tilde{P}) - \delta_0 \le \inf_x F(x, \tilde{P}) \le F(y^*, \tilde{P}),$$

and by virtue of (4.2) and (4.5), we have

$$|F(y^*, P) - F(y^*, \tilde{P})| \le \delta_1,$$
$$|F(x^*, P) - F(\tilde{x}^*, \tilde{P})| \le \delta_2.$$

This implies

$$\begin{aligned} F(y^*, P) &\ge F(y^*, \tilde{P}) - \delta_1 \\ &\ge F(\tilde{x}^*, \tilde{P}) - \delta_0 - \delta_1 \\ &\ge F(x^*, P) - \delta_0 - \delta_1 - \delta_2. \end{aligned}$$ □

COROLLARY 4.5. *Suppose that Assumption* 3 *holds, and*

$$\bar{d}_W(P, \tilde{P}) \le (\varepsilon_1, \dots, \varepsilon_T),$$

where P is supported by Ξ, and $\tilde{P}$ is supported by the finite set $\tilde{\Xi}$, and the approximation has been chosen so that the values of $\delta_1 = \delta_1(\varepsilon)$ and $\delta_2 = \delta_2(\varepsilon)$ are given. If $\tilde{x}^*$ is the solution of the approximate problem (4.1)*, then its extension x^* is a $(\delta_1 + \delta_2)$-solution of the basic problem* (3.3).

Proof. Choosing $\delta_0 = 0$ and applying Proposition 4.4 we get the statement immediately. □

5. Examples.

5.1. Multistage portfolio optimization. Consider an investor, who has initial capital C and wants to invest in m different assets. The price of one unit of asset i at time t, $t = 1, \dots, T$, is the random quantity $\xi_{t,i}$. At starting time 0, the prices $\xi_{0,i}$ are deterministic.

Let $\xi_t = (\xi_{t,1}, \dots, \xi_{t,m})^\top$ be the vector price process. The optimization problem is to maximize the acceptability of the final wealth under the self-financing constraint, i.e.,

$$\begin{array}{lll} \text{maximize in } x & \mathcal{A}[x_{T-1}^\top \xi_T] & \\ \text{subject to} & x_0^\top \xi_0 = C, & \\ & x_{t-1}^\top \xi_t = x_t^\top \xi_t, & t = 1, \dots, T-1, \end{array} \tag{5.1}$$

where $\mathcal{A}$ denotes some acceptability functional (see [1], [18]). By introducing the wealth w_t at time t as

$$w_t = x_{t-1}^\top \xi_t,$$

and defining the state as

$$\zeta_t = \begin{pmatrix} w_t \\ \xi_t \end{pmatrix},$$

one gets the dynamics

$$\begin{pmatrix} w_t \\ \xi_t \end{pmatrix} = \begin{pmatrix} x_{t-1}^\top \xi_t \\ \xi_t \end{pmatrix}, \qquad t = 1, \dots, T.$$

The constraint sets are

$$\mathcal{X}_t(\zeta_t) = \{x_t \,:\, x_t^\top \xi_t = w_t\}, \qquad t = 1, \dots, T.$$

Under the realistic assumption that the returns are bounded from above and from below,

$$0 < a \le \xi_{t,i} \le b < \infty,$$

this dynamics is Lipschitz in the sense of Assumption 2, since x must be bounded due to the initial budget constraint. In addition, the constraint sets are Lipschitz; see the second example in section 2.2.

The tree approximation of (5.1) is given by

$$\begin{array}{ll} \text{maximize in } x & \mathcal{A}[x_{T-1}^\top \tilde{\xi}_T] \\ \text{subject to} & x_0^\top \tilde{\xi}_0 = C, \\ & x_{t-1}^\top \tilde{\xi}_t = x_t^\top \tilde{\xi}_t, \qquad t = 1, \dots, T-1. \end{array}$$

Probability functionals which are Lipschitz w.r.t. the transportation distance in $\mathbb{R}$ include the expectation, the mean absolute deviation, distortion functionals (and therefore the average value-at-risk as a special case), and linear combinations thereof [17].

As illustration, let us consider the average value-at-risk corrected expectation [24]

$$\mathcal{A}[z] := \mathbb{E}[z] - \beta \mathbb{A}V@R_\alpha(z), \qquad 0 \le \beta \le 1,$$

where $\mathbb{A}V@R_\alpha(z)$ is defined by (A.1).

By

$$\left| a - \frac{1}{\alpha}\mathbb{E}([z-a]^-) - a + \frac{1}{\alpha}\mathbb{E}[[\tilde{z}-a]^-] \right| \le \frac{1}{\alpha}\mathbb{E}[|z-\tilde{z}|]$$

one sees that $\mathbb{A}V@R_\alpha$ is Lipschitz w.r.t. the transportation metric. The Lipschitz constant of $\mathbb{E}[z] - \beta\,\mathbb{A}V@R_\alpha(z)$ is $1 + \beta/\alpha$.

5.2. Multistage inventory control problem. We consider a generalization of the well-known newsboy problem (see, e.g., [12]) to a multiperiod setting. The multiperiod inventory model allows for storing the unsold merchandise, while the newsboy keeps no unsold copies, as they are worthless the next day.

Suppose that the demand at times $t = 1, \dots, T$ is given by a random process $\xi_1, \dots, \xi_T$. The regular orders are to be placed one period ahead. The order cost per unit ordered is one. Let I_t be the inventory level right after all sales have been effectuated at time t. If a stock-out O_t occurs, i.e., if the demand exceeds the inventory

plus the arriving order, the demand is satisfied by a rapid order. The rapid order cost per unit ordered is $r_t > 1$. The unsold goods are stored, but a fraction $(1 - q_t)$ is a storage loss of the period t; i.e., the inventory volume at the beginning of the period $(t+1)$ is $q_t I_t$. The selling price at time t is $s_t > 1$. Notice that all prices may change from period to period. The decision x_t is the order size at time t, $t = 0, \ldots, T-1$.

A closer look at the inventory volume and shortage shows that $I_0 = O_0 = 0$, and for $t = 1, \ldots, T$,

$$I_t = \max(q_{t-1}I_{t-1} + x_{t-1} - \xi_t, 0) = [q_{t-1}I_{t-1} + x_{t-1} - \xi_t]^+$$

and

$$O_t = \max(-(q_{t-1}I_{t-1} + x_{t-1} - \xi_t), 0) = [q_{t-1}I_{t-1} + x_{t-1} - \xi_t]^-.$$

For $t = 1, \ldots, T$, these two equations can be merged into one:

$$q_{t-1}I_{t-1} + x_{t-1} - \xi_t = I_t - O_t, \qquad I_t \geq 0,\ O_t \geq 0. \tag{5.2}$$

The profit function is

$$F(x, P) = q_T I_T + \sum_{t=1}^{T} \left(s_t \xi_t - x_{t-1} - r_t O_t\right),$$

and the optimization problem is to maximize the expected profit, i.e.,

$$\begin{array}{ll} \text{maximize in } x & q_T I_T + \displaystyle\sum_{t=1}^{T} \mathbb{E}[s_t \xi_t - x_{t-1} - r_t O_t] \\ \text{subject to} & x_t \text{ is } \mathcal{F}_t\text{-measurable, and (5.2) holds for } t = 1, \ldots, T. \end{array}$$

Notice that $\sum_{t=1}^{T} s_t \xi_t$ does not depend on the decision x and can be removed from the optimization problem.

As usual, x^* is the optimal solution, and denote by $v(x^*)$ the optimal value of

$$\begin{array}{ll} \text{maximize in } x & q_T I_T + \displaystyle\sum_{t=1}^{T} \mathbb{E}[-x_{t-1} - r_t O_t] \\ \text{subject to} & x_t \text{ is } \mathcal{F}_t\text{-measurable, and (5.2) holds for } t = 1, \ldots, T. \end{array} \tag{5.3}$$

Since (5.3) is linear in the decision variables $x_0, x_1, \ldots, x_{T-1}$, it has a dual formulation. In order to obtain it, let us form the Lagrangian $L(x, I, O, \lambda)$, for some $\lambda = (\lambda_1, \ldots, \lambda_T) \in L_\infty$, and $I = (I_1, \ldots, I_T)$, $O = (O_1, \ldots, O_T)$:

$$\begin{aligned} & L(x, I, O, \lambda) \\ &= \mathbb{E}\left[q_t I_T + \sum_{t=1}^{T}(-x_{t-1} - r_t O_t)\right] - \mathbb{E}\left[\sum_{t=1}^{T} \lambda_t (I_t - O_t - q_{t-1}I_{t-1} - x_{t-1} + \xi_t)\right] \\ &= \sum_{t=1}^{T} \mathbb{E}[x_{t-1}(\lambda_t - 1)] + \sum_{t=1}^{T} \mathbb{E}[O_t(\lambda_t - r_t)] \\ &\quad + \mathbb{E}[I_T(q_T - \lambda_T)] + \sum_{t=1}^{T-1} \mathbb{E}[(I_t(q_t\lambda_{t+1} - \lambda_t)] - \sum_{t=1}^{T} \mathbb{E}[\xi_t \lambda_t] \end{aligned}$$

$$= \sum_{t=1}^{T} \mathbb{E}[x_{t-1}(\mathbb{E}[\lambda_t|\mathcal{F}_{t-1}] - 1)] + \sum_{t=1}^{T} \mathbb{E}[O_t(\lambda_t - r_t)]$$
$$+ \mathbb{E}[I_T(q_T - \lambda_T)] + \sum_{t=1}^{T-1} \mathbb{E}[I_t(q_t\mathbb{E}[\lambda_{t+1}|\mathcal{F}_t] - \lambda_t)] + \sum_{t=1}^{T} \mathbb{E}[(-\xi_t)\lambda_t].$$

Only if $\mathbb{E}[\lambda_t|\mathcal{F}_{t-1}] = 1$, and $q_t \le \lambda_t \le r_t$, a.s. for each $t = 1, \dots, T$, is the dual problem finite. Thus, the dual formulation of (5.3) reduces to the following:

$$\begin{array}{ll} \text{minimize in } \lambda & \displaystyle\sum_{t=1}^{T} \mathbb{E}[(-\xi_t)\lambda_t] \\ \text{subject to} & \lambda_t, I_T, O_t \text{ are } \mathcal{F}_t\text{-measurable,} \\ & q_t \le \lambda_t \le r_t, \;\; I_t \ge 0, \;\; O_t \ge 0, \;\; \text{and} \\ & \mathbb{E}[\lambda_t|\mathcal{F}_t] = 1, \;\; \text{for } t = 1, \dots, T, \;\; \text{a.s.} \end{array} \tag{5.4}$$

Now recall the dual representation of $\mathbb{A}V@R_\alpha(\xi)$ given by (A.2). Setting, for $t = 1, \dots, T$,

$$Y_t = \frac{\lambda_t - q_t}{1 - q_t}$$

yields

$$\mathbb{E}[Y_t|\mathcal{F}_t] = 1 \quad \text{and} \quad 0 \le Y_t \le \frac{r_t - q_t}{1 - q_t}.$$

So, for $\alpha_t = \frac{1-q_t}{r_t-q_t}$, according to (A.5), the objective to be minimized in (5.4) becomes

$$\sum_{i=1}^{T} \mathbb{E}\left[\mathbb{A}V@R_{\alpha_t}(\xi_t)\right] = \sum_{i=1}^{T} \Big(q_t\mathbb{E}[-\xi_t] + (1 - q_t)\mathbb{E}[\mathbb{A}V@R_{\alpha_t}(-\xi_t|\mathcal{F}_{t-1})] \Big).$$

Using the identity (A.3) with $\beta_t = 1 - \alpha_t = \frac{r_t-1}{r_t-q_t}$, we write the objective as

$$\sum_{i=1}^{T} \Big(r_t\mathbb{E}[-\xi_t] + (r_t - 1)\mathbb{E}[\mathbb{A}V@R_{\beta_t}(\xi_t|\mathcal{F}_t)] \Big). \tag{5.5}$$

The optimal solution of the given problem, for $\mathbb{V}@R_{\beta_{t+1}}(\xi_{t+1}|\mathcal{F}_t) = V_t$, is

$$x_t^* = \mathbb{V}@R_{\beta_{t+1}}(\xi_{t+1}|\mathcal{F}_t) - q_tI_t = V_t - q_tI_t, \qquad t = 0, \dots, T-1. \tag{5.6}$$

Indeed, if we insert (5.6) into (5.5), in view of (A.4) we get

$$v(x^*) = \sum_{i=1}^{T} \Big(r_t\mathbb{E}[-\xi_t] + (r_t - 1)\mathbb{E}[V_{t-1}] - (r_t - q_t)\mathbb{E}[[V_{t-1} - \xi_t]^+] \Big).$$

On the other hand, inserting the solutions (5.6) into the constraints of (5.3), we have that, for $t = 1, \dots, T$,

$$I_t = [V_{t-1} - \xi_t]^+,$$
$$O_t = [V_{t-1} - \xi_t]^-,$$

and (5.2) is

$$V_{t-1} - \xi_t = I_t - O_t, \qquad I_t \geq 0,\ O_t \geq 0.$$

The value of the objective in (5.3) for this choice of x becomes

$$\begin{aligned}
&q_T I_T + \sum_{t=1}^{T} \mathbb{E}[-x_{t-1} - r_t O_t] \\
&= \mathbb{E}\left[q_T I_t + \sum_{t=1}^{T} (-V_{t-1} - q_{t-1} I_{t-1} - r_t O_t) \right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} (q_t I_t - V_{t-1} - r_t (I_t - V_{t-1} + \xi_t)) \right] \\
&= \sum_{i=1}^{T} \Big(r_t \mathbb{E}[-\xi_t] + (r_t - 1)\mathbb{E}[V_{t-1}] - (r_t - q_t)\mathbb{E}[[V_{t-1} - \xi_t]^+] \Big) \\
&= v(x^*).
\end{aligned}$$

Thus, we have shown that (5.6) is really the solution of the given problem.

The state of the system at time $t = 1, \ldots, T$ is

$$\zeta_t = (\xi_1, \ldots, \xi_t, I_t, O_t, x_t).$$

The mapping

$$\begin{gathered}
\zeta_t \mapsto x_t^*, \ \text{i.e.,} \\
(\xi_1, \ldots, \xi_t, I_t) \mapsto \mathbb{V}@R_{\beta_t}(\xi_{t+1} | \mathcal{F}_t) - q_t I_t,
\end{gathered}$$

is Lipschitz if the mapping

$$(\xi_1, \ldots, \xi_t) \mapsto \mathbb{V}@R_{\beta_t}(\xi_{t+1} | \mathcal{F}_t) \tag{5.7}$$

is Lipschitz. This is true in many cases, e.g., for vector autoregressive processes.

Let $G(v|u_1, \ldots, u_{t-1})$ be the conditional distribution function of ξ_t given the past $\xi^{t-1} = u^{t-1}$. If

$$(u_1, \ldots, u_t) \mapsto G^{-1}(v|u_1, \ldots, u_t)$$

is Lipschitz for all v, then (5.7) is also Lipschitz.

In order to show how the Lipschitz constants in the inventory model behave, assume that the demand process $\xi = (\xi_1, \ldots, \xi_T)$ is normally distributed and follows an additive recursion. To this end, let

$$\xi_0 \sim \mathcal{N}(\mu_0, \sigma_0^2), \qquad \xi_t = b\xi_{t-1} + \epsilon_t \text{ with } \epsilon_t \sim \mathcal{N}(\mu, \sigma^2), \qquad t = 1, \ldots, T,$$

where $\mu = \mu_0(1 - b)$, $\sigma = \sigma_0(1 - b^2)$, and ξ_t and ϵ_t are independent. Under these assumptions, ξ is a stationary Gaussian Markov process.

The state of the system may be reduced to

$$\zeta_t = (\xi_t, I_t, O_t, x_t), \qquad t = 1, \ldots, T. \tag{5.8}$$

The conditional average value-at-risk is

$$\begin{aligned}\mathbb{A}V@R_\beta(\xi_t|\mathcal{F}_{t-1}) &= b\xi_{t-1} + \mathbb{A}V@R_\beta(\epsilon_t)\\ &= b\xi_{t-1} + \mu - \frac{1}{\beta\sqrt{2\pi}}\exp\left(\frac{1}{2}\Big(\Phi^{-1}(\min(\beta, 1-\beta))\Big)^2\right).\end{aligned}$$

The solution of (5.6) becomes

$$x_t^* = b\xi_{t-1} + \mu - \frac{1}{\beta_t\sqrt{2\pi}}\exp\left(\frac{1}{2}\Big(\Phi^{-1}(\min(\beta_t, 1-\beta_t))\Big)^2\right) - q_{t-1}I_{t-1}.$$

To calculate the constants K_t from Assumption 2, we need the conditional distribution of ξ_t given $\xi_{t-1} = u_{t-1}$. It holds that

$$(\xi_t|\mathcal{F}_{t-1}) = (\xi_t|\xi_{t-1} = u) \sim \mathcal{N}(\mu + bu, \sigma^2).$$

Thus,

$$d_W(P_t(\cdot|u), P_t(\cdot|v)) \le bd(u, v),$$

and $K_t = b$, for each $t = 1, \dots, T$.

For constants L_t, M_t, N_t from Assumption 3, observe the state ζ_t given by (5.8), and recall the transition function g_t defined by (3.1). Then

$$\begin{aligned}\zeta_{t+1} = (\xi_{t+1}, I_{t+1}, O_{t+1}, x_{t+1}) &= g_{t+1}((\zeta_t, I_t, O_t, x_t), x_t, \xi_{t+1})\\ &= g_{t+1}(g_t((\zeta_{t-1}, I_{t-1}, O_{t-1}, x_{t-1}), x_{t-1}, \xi_t), I_t, x_t, \xi_{t+1}),\end{aligned}$$

and so on. It follows that

$$\begin{aligned}&d(g_t(\zeta_{t-1}, x_{t-1}, \xi_t), g_t(\bar\zeta_{t-1}, \bar x_{t-1}, \bar\xi_t))\\ &= d\left([q_{t-1}I_{t-1} + x_{t-1} - \xi_t]^+ - [q_{t-1}\bar I_{t-1} + \bar x_{t-1} - \bar\xi_t]^+\right)\\ &\le \max(0, 1, 0, 0)\left(d(I_{t-1}, \bar I_{t-1}) + d(x_{t-1}, \bar x_{t-1}) + d(\xi_t, \bar\xi_t)\right),\end{aligned}$$

and $L_t = 1$. Since $\mathcal{X}_t(z) = \mathbb{R}^+$ for all z, we have that $d_{P,r}(\mathcal{X}_t(z), \mathcal{X}_t(\bar z)) = 0$, and trivially $M_t = 0$. For N_t, we see that

$$|\mathbb{F}_t(\zeta_t) - \mathbb{F}_t(\bar\zeta_t)| \le \max(1, 1, r_t)d_W(\zeta_t, \bar\zeta_t),$$

and so $N_t = r_t$.

The Lipschitz constant of the solution, as well as of the extension, is $Q_t = Q_t^e = 1$ for $t = 0, \dots, T-1$.

Eventually, we want to calculate constants δ_1, resp., δ_2, as obtained in Proposition 4.2, resp., Proposition 4.3. In this setting, we have that

$$\begin{aligned}\bar\varepsilon_s &= \sum_{i=1}^{s}\varepsilon_i b^{s-i}, \qquad s = 1, \dots, T,\\ D_{t,s} = D_{t,s}^e &= t - s + 1, \qquad s = 1, \dots, t-1,\end{aligned}$$

and

$$\delta_1 = \delta_2 = \sum_{s=1}^{T}\left(\sum_{i=1}^{s}\varepsilon_i b^{s-i}\sum_{t=s}^{T}(t-s+1)r_t\right).$$

Appendix. Auxiliary results. The following lemma is a generalization of [23, Chapter 7J, Exercise 7.67].

LEMMA A.1. *Let $\mathbb{B}(r)$ be as in section 2.2 and let $A, B \subseteq \mathbb{B}(r)$ be closed, convex sets for some $r > 0$. Let y be some point in $\mathbb{B}(r)$. Then*

$$\|proj_A(y) - proj_B(y)\| \leq 2\sqrt{r\, d_r(A,B)},$$

where

$$d_r(A,B) = \sup_{x \in \mathbb{B}(r)} |d(x,A) - d(x,B)|.$$

Proof. Define the half spaces

$$H_A = \{x : (x - proj_A(y))^\top (y - proj_A(y)) \leq 0\} \text{ and}$$
$$H_B = \{x : (x - proj_B(y))^\top (y - proj_B(y)) \leq 0\}.$$

Then $A \subseteq H_A$ and $B \subseteq H_B$. Assume without loss of generality that $\|y - proj_A(y)\| \leq \|y - proj_B(y)\|$.

Let $y' = proj_{H_B}(proj_A(y))$ and let y'' be the projection of y on the line connecting $proj_A(y)$ and $proj_B(y)$. Then, because $\|y - proj_A(y)\| \leq \|y - proj_B(y)\|$, one has that

$$\|y'' - proj_B(y)\| \geq \frac{1}{2}\|proj_A(y) - proj_B(y)\|.$$

Notice that

$$d_r(A,B) \geq d(proj_A(y), H_B) = \|proj_A(y) - y'\|,$$

since $proj_A(y) \in A$, and $B \subseteq H_B$. The five points $(y, proj_A(y), proj_B(y), y', y'')$ lie in one hyperplane. By geometrical consideration, using the similarity of triangles, one has

$$\frac{\|proj_A(y) - y'\|}{\|proj_A(y) - proj_B(y)\|} = \frac{\|y'' - proj_B(y)\|}{\|y - proj_B(y)\|} \geq \frac{\|proj_A(y) - proj_B(y)\|}{2\,\|y - proj_B(y)\|}.$$

Hence, using $\|y - proj_B(y)\| \leq 2r$, one gets

$$\|proj_A(y) - proj_B(y)\|^2 \leq 2\|y - proj_B(y)\|\,\|proj_A(y) - y'\| \leq 4r\, d_r(A,B). \qquad \square$$

For a real-valued Lipschitz function in finite-dimensional spaces, extension from an arbitrary subset is possible.

THEOREM A.2 (extension theorem). *Let $(\tilde{\Xi}, d)$ be any metric space, $\tilde{\Xi}$ any subset of Ξ, and x any $\mathbb{R}^m$-valued Lipschitz function on $\tilde{\Xi}$. Then x can be extended on Ξ without increasing the Lipschitz modulus.*

Proof. See [5, Theorem 6.1.1]. $\square$

In [8] one can read more about the extension of Lipschitz functions in the infinite-dimensional setting.

LEMMA A.3. *Let $x'(u)$ be a Q-Lipschitz function with values in $\mathbb{B}(r) \subseteq \mathbb{R}^m$ and let $\mathcal{X}(u)$ be a convex-valued multifunction which is M-Lipschitz w.r.t. the projection distance $d_{P,r}$. (The case $r = \infty$ is not excluded.) Then the convex projection $x(u) = proj_{\mathcal{X}(u)} x'(u)$ is $(Q + M)$-Lipschitz.*

Proof.

$$\begin{aligned} d(x(u),x(v)) &\le d(proj_{\mathcal{X}(u)}x'(u), proj_{\mathcal{X}(u)}x'(v)) + d(proj_{\mathcal{X}(u)}x'(v), proj_{\mathcal{X}(v)}x'(v)) \\ &\le d(x'(u),x'(v)) + d_{P,r}(\mathcal{X}(u),\mathcal{X}(v)) \\ &\le Qd(u,v) + Md(u,v) = (Q+M)d(u,v). \qquad \square \end{aligned}$$

LEMMA A.4. *The average value-at-risk is given by the following expressions:*

$$\begin{aligned} \mathbb{A}V@R_\alpha(\xi) &= \frac{1}{\alpha}\int_0^\alpha G^{-1}(u)\,du \\ &= \max_{a\in\mathbb{R}}\left(a - \frac{1}{\alpha}\mathbb{E}[[\xi-a]^-]\right) \qquad \text{(A.1)} \\ &= \min_Y\left\{\mathbb{E}[\xi Y] \,:\, 0\le Y\le\frac{1}{\alpha},\ \mathbb{E}[Y]=1\right\}, \qquad \text{(A.2)} \end{aligned}$$

where $G(u)=\mathbb{P}(\xi\le u)$. *If* $\mathbb{V}@R_\alpha(\xi)=G^{-1}(\alpha)$, *the following identities hold:*

$$\mathbb{A}V@R_\alpha(-Y) = \frac{1-\alpha}{\alpha}\mathbb{A}V@R_{1-\alpha}(Y) - \frac{1}{\alpha}\mathbb{E}[Y] \qquad \text{(A.3)}$$

and

$$\mathbb{A}V@R_\alpha(Y) = \mathbb{V}@R_\alpha(Y) - \frac{1}{\alpha}\mathbb{E}[[\mathbb{V}@R_\alpha(Y)-Y]^+]. \qquad \text{(A.4)}$$

The expected conditional average value-at-risk given the filtration $\mathcal{F}$ *is defined by*

$$\mathbb{E}[\mathbb{A}V@R_\alpha(\xi|\mathcal{F})] = \min_Y\left\{\mathbb{E}[\xi Y] \,:\, 0\le Y\le\frac{1}{\alpha},\ \mathbb{E}[Y|\mathcal{F}]=1\right\}. \qquad \text{(A.5)}$$

Proof. See [18] and [19]. $\square$

REFERENCES

[1] P. ARTZNER, F. DELBAEN, J.-M. EBER, AND D. HEATH, *Coherent measures of risk*, Math. Finance, 9 (1999), pp. 203–228.

[2] Y. S. CHOW AND U. GRENANDER, *A sieve method for the spectral density*, Ann. Statist., 13 (1985), pp. 998–1010.

[3] M. DELECROIX, M. SIMIONI, AND C. THOMAS-AGNAN, *Functional estimation under shape constraints*, J. Nonparametr. Statist., 6 (1996), pp. 69–89.

[4] R. DOBRUSHIN, *Central limit theorem for non-stationary Markov chains.* I, Teor. Veroyatnost. i Primenen., 1 (1956), pp. 72–89.

[5] R. M. DUDLEY, *Real Analysis and Probability*, Wadsworth Brooks/Cole Math. Ser., Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1989.

[6] R. DURRETT, *Probability: Theory and Examples*, 2nd ed., Duxbury Press, Belmont, CA, 2004.

[7] S. EFROMOVICH, *Nonparametric Curve Estimation. Methods, Theory, and Applications*, Springer Ser. Statist., Springer-Verlag, New York, 1999.

[8] L. GEHÉR, *Über Fortsetzungs- und Approximationsprobleme für stetige Abbildungen von metrischen Räumen*, Acta Sci. Math. Szeged, 20 (1959), pp. 48–66.

[9] H. HEITSCH, W. RÖMISCH, AND C. STRUGAREK, *Stability of multistage stochastic programs*, SIAM J. Optim., 17 (2006), pp. 511–525.

[10] R. HOCHREITER AND G. C. PFLUG, *Financial scenario generation for stochastic multi-stage decision processes as facility location problems*, Ann. Oper. Res., 152 (2007), pp. 257–272.

[11] R. S. LIPTSER AND A. N. SHIRYAYEV, *Statistics of Random Processes.* II. *Applications*, Appl. Math. 6, Springer-Verlag, New York, 1978.

[12] S. NAHMIAS, *Production and Operations Analysis*, 5th ed., McGraw–Hill, New York, 2005.

[13] P. Olsen, *Discretizations of multistage stochastic programming problems*, in Stochastic Systems: Modeling, Identification and Optimization, II (Lexington, KY, 1975), Math. Programming Stud. 6, North–Holland, Amsterdam, 1976, pp. 111–124.
[14] T. Pennanen, *Epi-convergent discretizations of multistage stochastic programs*, Math. Oper. Res., 30 (2005), pp. 245–256.
[15] G. C. Pflug, *Optimization of Stochastic Models. The Interface between Simulation and Optimization*, Kluwer Internat. Ser. Engrg. Comput. Sci. 373, Kluwer Academic Publishers, Boston, MA, 1996.
[16] G. C. Pflug, *Scenario tree generation for multiperiod financial optimization by optimal discretization*, Math. Program., 89 (2001), pp. 251–271.
[17] G. C. Pflug, *On distortion functionals*, Statist. Decisions, 24 (2006), pp. 45–60.
[18] G. C. Pflug, *Subdifferential representations of risk measures*, Math. Program., 108 (2006), pp. 339–354.
[19] G. C. Pflug and W. Römisch, *Modeling, Measuring and Managing Risk*, World Scientific, Singapore, 2007.
[20] S. T. Rachev, *Probability Metrics and the Stability of Stochastic Models*, Wiley Ser. Probab. Math. Statist. Appl. Probab. Statist., John Wiley and Sons, Chichester, UK, 1991.
[21] T. Robertson and F. T. Wright, *Consistency in generalized isotonic regression*, Ann. Statist., 3 (1975), pp. 350–362.
[22] R. T. Rockafellar and R. J. B. Wets, *Continuous versus measurable recourse in N-stage stochastic programming*, J. Math. Anal. Appl., 48 (1974), pp. 836–859.
[23] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Grundlehren Math. Wiss. 317, Springer-Verlag, Berlin, 2004.
[24] S. Uryasev and R. T. Rockafellar, *Conditional value-at-risk: Optimization approach*, in Stochastic Optimization: Algorithms and Applications (Gainesville, FL, 2000), Appl. Optim. 54, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001, pp. 411–435.
[25] S. S. Vallander, *Calculations of the Vasserštein distance between probability distributions on the line*, Teor. Verojatnost. i Primenen., 18 (1973), pp. 824–827.
[26] C. Villani, *Topics in Optimal Transportation*, Grad. Stud. Math. 58, AMS, Providence, RI, 2003.
[27] V. Zolotarev, *Probability metrics*, Theory Probab. Appl., 28 (1983), pp. 278–302.

SIAM J. OPTIM.
Vol. 18, No. 3, pp. 1106–1127

NONSMOOTH ANALYSIS OF LORENTZ INVARIANT FUNCTIONS*

HRISTO S. SENDOV†

Abstract. A real valued function $g(x,t)$ on $\mathbb{R}^n \times \mathbb{R}$ is called a *Lorentz invariant* if $g(x,t) = g(Ux,t)$ for all $n \times n$ orthogonal matrices U and all (x,t) in the domain of g. In other words, g is invariant under the linear orthogonal transformations preserving the *Lorentz cone*: $\{(x,t) \in \mathbb{R}^n \times \mathbb{R} \,|\, t \geq \|x\|\}$. It is easy to see that every Lorentz invariant function can be decomposed as $g = f \circ \beta$, where $f : \mathbb{R}^2 \to \mathbb{R}$ is a symmetric function and β is the root map of the *hyperbolic polynomial* $p(x,t) = t^2 - x_1^2 - \cdots - x_n^2$. We investigate a variety of important variational and nonsmooth properties of g and characterize them in terms of the symmetric function f.

Key words. nonsmooth analysis, convex analysis, hyperbolic polynomials, Lorentz cone, second-order cone, Clarke subdifferential, regular subdifferential, limiting subdifferential, proximal subdifferential, lower semicontinuous

AMS subject classifications. Primary, 49J52, 58C20; Secondary, 58C25, 58E30

DOI. 10.1137/060658370

1. Introduction and notation.

Denote the set of all orthogonal $n \times n$ matrices by $O(n)$. Let the function $g(x,t)$ be defined on an open subset of $\mathbb{R}^n \times \mathbb{R}$, taking values in $\mathbb{R}$. The inner product of two vectors, (x,t) and (y,r), in $\mathbb{R}^n \times \mathbb{R}$ is $\langle (x,t),(y,r),\rangle = x^T y + tr$. Throughout the entire paper we assume that

$$g(Ux,t) = g(x,t) \quad \text{for all } U \in O(n), \tag{1.1}$$

and all (x,t) in the domain of g. We call a function g with property (1.1) *Lorentz invariant* because it is invariant under the linear orthogonal transformations preserving the *Lorentz cone* $\{(x,t) \in \mathbb{R}^n \times \mathbb{R} \,|\, t \geq \|x\|\}$. A set $\Omega \subseteq \mathbb{R}^n \times \mathbb{R}$ is called *Lorentz invariant* if $(x,t) \in \Omega$ implies that $(Ux,t) \in \Omega$ for every $U \in O(n)$. Define the map

$$\beta(x,t) : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^2,$$
$$\beta(x,t) = \frac{1}{\sqrt{2}}(t + \|x\|, t - \|x\|).$$

The rationale behind the map β is the following. Consider the polynomial $p(x,t) : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ defined by $p(x,t) = t^2 - x_1^2 - \cdots - x_n^2$ and let $d := (0,\ldots,0,\sqrt{2}) \in \mathbb{R}^n \times \mathbb{R}$. Then, the coordinates of $\beta(x,t)$ are the roots of the polynomial $\lambda \mapsto P((x,t) - \lambda d)$. In general, a homogeneous polynomial $p(x) : \mathbb{R}^n \to \mathbb{R}$ with degree of homogeneity m, for which there is a direction $d \in \mathbb{R}^n$, $p(d) \neq 0$, such that $\lambda \mapsto p(x - \lambda d)$ has m real roots for every $x \in \mathbb{R}^n$, is called *hyperbolic*. In 1997, Güler [6] pointed out the relevance of these polynomials for optimization. Further information and developments can be found in [2], [13], [12], [18].

Let the function $f(a,b)$ be defined on an open subset of $\mathbb{R}^2$ and assume that it is *symmetric*, that is, $f(a,b) = f(b,a)$ for all (a,b) in its domain. Necessarily, the

*Received by the editors April 27, 2006; accepted for publication (in revised form) July 11, 2007; published electronically October 4, 2007. The author's research was supported by NSERC.
http://www.siam.org/journals/siopt/18-3/65837.html
†Department of Statistical and Actuarial Sciences, University of Western Ontario, 1151 Richmond Street North, Western Science Centre, Room 262, London, ON, Canada N6A 5B7 (hssendov@stats.uwo.ca).

domain of f is a *symmetric subset* of $\mathbb{R}^2$, that is, $(a,b) \in A \Rightarrow (b,a) \in A$. The following easy lemma establishes the connection between g, β, and f.

LEMMA 1.1 (Lorentz invariant functions). *The following two properties of a function $g : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ are equivalent:*

(i) *g is Lorentz invariant;*

(ii) *$g = f \circ \beta$ for some symmetric function $f : \mathbb{R}^2 \to \mathbb{R}$.*

If $g = f \circ \beta$, we say that f is the symmetric function corresponding to g. This correspondence is one-to-one, and given g the corresponding symmetric function is

$$f(a,b) = g\left(\frac{a-b}{\sqrt{2}}, 0, \ldots, 0, \frac{a+b}{\sqrt{2}}\right). \tag{1.2}$$

That (1.2) defines a symmetric function in (a,b) is guaranteed by (1.1).

The goal of this paper is to establish a variety of important variational and nonsmooth optimization properties of the function $g = f \circ \beta$ and how they arise from the corresponding properties of f. By deriving a wide range of nonsmooth formulae we hope this work will be a useful reference source. This work completes the similar investigations of spectral functions [8], [9], [11], [7] and singular value functions [10], [14], [15]. Optimization problems over the Lorentz cone, also known as the second-order cone, have a wide range of applications; see, for example, [16]. With the development of the nonsmooth Newton method and various smoothing techniques, the nonsmooth properties of functions associated with the Lorentz cone have been of interest lately. For example, the strong semismoothness of the projection onto the Lorentz cone was established in [23, Proposition 4.3]. A formula for the Bouligand subdifferential of the projection onto the Lorentz cone is derived in [24, Lemma 14]. Our paper is based on results that first appeared in the author's Ph.D. dissertation [21].

We conclude this section with an elementary fact.

LEMMA 1.2. *The composition $f \circ \beta$ is lower semicontinuous if and only if f is lower semicontinuous.*

Throughout the entire work, the functions g, β, and f will have the properties described in this section.

2. Fenchel conjugation. For a function $F : \mathbb{R}^n \to (-\infty, +\infty]$, the *Fenchel conjugate* $F^* : \mathbb{R}^n \to [-\infty, +\infty]$ is the function

$$F^*(y) = \sup_{x \in \mathbb{R}^n} \{x^T y - F(x)\}.$$

It is well known that F^* is lower semicontinuous and convex [19]. In this section we prove the following formula.

PROPOSITION 2.1. *We always have*

$$(f \circ \beta)^* = f^* \circ \beta. \tag{2.1}$$

Proof. Let $y \neq 0$. In the third equality below, we use the fact that f is symmetric to see that the given supremum is the same as the supremum over the set $\{(a,b) \in \mathbb{R}^2 \mid a - b \geq 0\}$. From the definition we have

$$\begin{aligned}
(f \circ \beta)^*(y,r) &= \sup_{(x,t) \in \mathbb{R}^{n+1}} \{\langle (y,r), (x,t) \rangle - (f \circ \beta)(x,t)\} \\
&= \sup_{(a,b) \in \mathbb{R}^2} \sup_{\substack{(x,\,t) \ \text{s.t.} \\ t + \|x\| = a\sqrt{2} \\ t - \|x\| = b\sqrt{2}}} \{\langle (y,r), (x,t) \rangle - f(a,b)\} \\
&= \sup_{(a,b) \in \mathbb{R}^2} \left\{ \left\langle (y,r), \left(\frac{y}{\|y\|} \frac{a-b}{\sqrt{2}}, \frac{a+b}{\sqrt{2}} \right) \right\rangle - f(a,b) \right\}
\end{aligned}$$

$$
\begin{aligned}
&= \sup_{(a,b)\in\mathbb{R}^2} \left\{ \|y\| \frac{a-b}{\sqrt{2}} + r\frac{a+b}{\sqrt{2}} - f(a,b) \right\} \\
&= \sup_{(a,b)\in\mathbb{R}^2} \left\{ \left\langle \left(\frac{r+\|y\|}{\sqrt{2}}, \frac{r-\|y\|}{\sqrt{2}} \right), (a,b) \right\rangle - f(a,b) \right\} \\
&= (f^* \circ \beta)(y,r).
\end{aligned}
$$

The case $y = 0$ is easy. □

An alternative proof of this result uses Theorem 5.5 and the example in section 7.5 in [2], where the proposition has been generalized to the subclass of so-called *isometric* hyperbolic polynomials. In [1, Theorem 6.1] the proposition has been shown to hold for symmetric functions composed with the eigenvalues of the elements of formally real Jordan algebras.

3. Convexity and convex subdifferentials.

3.1. Convexity.

THEOREM 3.1. *The composition $f \circ \beta$ is convex and lower semicontinuous if and only if f is convex and lower semicontinuous.*

Proof. Suppose f is convex and lower semicontinuous. If $f \equiv +\infty$, then $f \circ \beta \equiv +\infty$ and the theorem is clear. Suppose f assumes some finite values. Then using the convexity, one can show that $f > -\infty$, and by [19, Theorem 12.2] we have $f^{**} = f$. Since f^* is symmetric, we use (2.1) in $f \circ \beta = f^{**} \circ \beta = (f^* \circ \beta)^*$ to conclude that $f \circ \beta$ is convex and lower semicontinuous. The opposite direction follows from (1.2) and Lemma 1.2. □

The proof of the above theorem can be also deduced from Theorem 3.9 and the example in section 7.5 in [2]. Even though the proof of our Theorem 3.1 is quite elegant, a direct approach removes the condition that f be lower semicontinuous.

THEOREM 3.2. *The composition $f \circ \beta$ is convex if and only if f is convex.*

Proof. If $f \circ \beta$ is convex, then f is by (1.2). Suppose now that f is convex with domain C. The domain of $f \circ \beta$ is $\beta^{-1}(C)$. Let (x,t), $(y,r) \in \beta^{-1}(C)$, and $\alpha \in [0,1]$. Since $(t+\|x\|, t-\|x\|), (r+\|y\|, r-\|y\|) \in \sqrt{2}C$, and C is symmetric and convex, we find that the points

$$
\begin{aligned}
&(\alpha t + (1-\alpha)r + \alpha\|x\| + (1-\alpha)\|y\|, \alpha t + (1-\alpha)r - \alpha\|x\| - (1-\alpha)\|y\|), \\
&(\alpha t + (1-\alpha)r - \alpha\|x\| - (1-\alpha)\|y\|, \alpha t + (1-\alpha)r + \alpha\|x\| + (1-\alpha)\|y\|)
\end{aligned}
$$

are both in $\sqrt{2}C$. Denote the first displayed point by $a\sqrt{2}$ and the second by $b\sqrt{2}$. Since

$$
-\alpha\|x\| - (1-\alpha)\|y\| \le \|\alpha x + (1-\alpha)y\| \le \alpha\|x\| + (1-\alpha)\|y\|,
$$

there is a $\beta \in [0,1]$ such that for the point

$$
c\sqrt{2} := (\alpha t + (1-\alpha)r + \|\alpha x + (1-\alpha)y\|, \alpha t + (1-\alpha)r - \|\alpha x + (1-\alpha)y\|)
$$

we have $c = \beta a + (1-\beta)b \in C$. Thus,

$$
\begin{aligned}
f(c) &\le \beta f(a) + (1-\beta) f(b) = f(a) \\
&\le \alpha f((t+\|x\|, t-\|x\|)/\sqrt{2}) + (1-\alpha) f((r+\|y\|, r-\|y\|)/\sqrt{2}),
\end{aligned}
$$

where we used the facts that $f(a) = f(b)$ and that f is convex. □

The proof of Theorem 3.2 shows the following property.

LEMMA 3.3. *Let $C \subseteq \mathbb{R}^2$ be a convex and symmetric set. Then*

$$\beta^{-1}(C) := \{(x,t) \in \mathbb{R}^n \times \mathbb{R} \,|\, \beta(x,t) \in C\}$$

is convex and Lorentz invariant.

3.2. Convex subdifferentials. Let $f : \mathbb{R}^2 \to (-\infty, +\infty]$ be convex. For every point (a,b) such that $f(a,b) < +\infty$, we define the *subdifferential* of f at (a,b) to be the set

$$\partial f(a,b) = \{(a',b') | f(c,d) - f(a,b) \geq \langle (a',b'), (c,d) - (a,b) \rangle \text{ for all } (c,d)\}.$$

It is easy to see that $f(a,b) + f^*(a',b') = \langle (a,b), (a',b') \rangle$ if and only if $(a',b') \in \partial f(a,b)$. The set $\partial f(a,b)$ is a singleton $\{(a',b')\}$ if and only if f is differentiable at the point (a,b) with gradient $\nabla f(a,b) = (a',b')$; see [19, Theorem 25.1].

The following result gives a formula for the subgradient of the composition $f \circ \beta$.

THEOREM 3.4. *Suppose $f : \mathbb{R}^2 \to (-\infty, +\infty]$ is convex and lower semicontinuous. Then $(y,r) \in \partial(f \circ \beta)(x,t)$ if and only if $\beta(y,r) \in \partial f(\beta(x,t))$ and $x^T y = \|x\|\|y\|$.*

Proof. Suppose first that $(y,r) \in \partial(f \circ \beta)(x,t)$. Then using formula (2.1) we get

$$\begin{aligned}
\|x\|\|y\| + rt \;\geq x^T y + rt &= \langle (y,r), (x,t) \rangle \\
&= (f \circ \beta)(x,t) + (f \circ \beta)^*(y,r) \\
&= (f \circ \beta)(x,t) + (f^* \circ \beta)(y,r) \\
&= f\left(\frac{t + \|x\|}{\sqrt{2}}, \frac{t - \|x\|}{\sqrt{2}} \right) + f^*\left(\frac{r + \|y\|}{\sqrt{2}}, \frac{r - \|y\|}{\sqrt{2}} \right) \\
&\geq \big((t + \|x\|)(r + \|y\|) + (t - \|x\|)(r - \|y\|) \big)/2 \\
&= \|x\|\|y\| + rt.
\end{aligned}$$

Thus, we have equalities everywhere: $\beta(y,r) \in \partial f(\beta(x,t))$ and $x^T y = \|x\|\|y\|$. In the other direction the proof is clear by reversing the steps above. ∎

For a generalization of this proposition to formally real Jordan algebras see [1, Corollary 6.2].

4. Differentiability. The partial derivatives of the function f with respect to its first and second argument are denoted by f_1' and f_2', respectively.

THEOREM 4.1. *The composition $f \circ \beta$ is differentiable at the point (x,t) if and only if f is differentiable at $\beta(x,t)$. In that case we have the formulae*

$$\nabla_x (f \circ \beta)(x,t) = \begin{cases} \frac{f_1'(\beta(x,t)) - f_2'(\beta(x,t))}{\sqrt{2}\|x\|} x & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

and

$$\frac{d}{dt}(f \circ \beta)(x,t) = \frac{1}{\sqrt{2}}(f_1'(\beta(x,t)) + f_2'(\beta(x,t))).$$

Proof. Suppose first that f is differentiable at the point $\beta(x,t)$. If $x \neq 0$, the theorem and the formulae are trivial and follow from the chain rule. So let us assume now that $x = 0$. Let $h = (\bar{h}, h_{n+1}) \in \mathbb{R}^n \times \mathbb{R}$ and

$$d := \big(0, \ldots, 0, (f_1'(\beta(x,t)) + f_2'(\beta(x,t)))/\sqrt{2}\big) \in \mathbb{R}^n \times \mathbb{R}.$$

Then

$$\lim_{h\to 0}\frac{|(f\circ\beta)((0,t)+(\bar{h},h_{n+1}))-(f\circ\beta)((0,t))-d^Th|}{\|h\|}$$
$$=\lim_{h\to 0}\frac{|f(\beta(\bar{h},t+h_{n+1}))-f(\beta(0,t))-h_{n+1}(f_1'(\beta(0,t))+f_2'(\beta(0,t)))/\sqrt{2}|}{\|h\|}.$$

The fact that f is differentiable at $\beta(0,t)=(t/\sqrt{2},t/\sqrt{2})$ gives

$$f(\beta(\bar{h},t+h_{n+1}))\sim f(\beta(0,t))+f_1'(\beta(0,t))\frac{h_{n+1}+\|\bar{h}\|}{\sqrt{2}}+f_2'(\beta(0,t))\frac{h_{n+1}-\|\bar{h}\|}{\sqrt{2}},$$

where $\sim$ indicates that the difference of both sides is of order $o(\|h\|)$. Using the fact that, for a symmetric function f, $f_1'(\beta(0,t))=f_2'(\beta(0,t))$ and substituting above we see that the limit is zero, that is, $\nabla(f\circ\beta)(0,t)=d$.

The proof in the other direction is easy using formula (1.2). □

THEOREM 4.2. *Let f be symmetric and defined on an open symmetric subset of $\mathbb{R}^2$. Then $f\circ\beta$ is continuously differentiable at the point (x,t) if and only if f is continuously differentiable at $\beta(x,t)$.*

Proof. Suppose that f is continuously differentiable at $\beta(x,t)$. The theorem is clear if $x\neq 0$. So suppose $x=0$. Let $\{(x^k,t^k)\}$ be a sequence of points in $\mathbb{R}^n\times\mathbb{R}$ approaching $(0,t)$. We need only prove that $\nabla(f\circ\beta)(x^k,t^k)$ approaches $\nabla(f\circ\beta)(0,t)$. We consider two cases. The general case easily follows by combining them.

Case 1. If $x^k=0$ for all k, then using the formula in Theorem 4.1 we obtain

$$\lim_{k\to\infty}\nabla(f\circ\beta)(0,t^k)=\lim_{k\to\infty}\left(0,\dots,0,\frac{1}{\sqrt{2}}\big(f_1'(\beta(0,t^k))+f_2'(\beta(0,t^k))\big)\right)$$
$$=\nabla(f\circ\beta)(0,t)$$

by the continuity of ∇f at $\beta(0,t)$.

Case 2. If $x^k\neq 0$ for all k, then using again the formula in Theorem 4.1 for the derivative with respect to t we obtain

$$\lim_{k\to\infty}(f\circ\beta)_t'(x^k,t^k)=\lim_{k\to\infty}\frac{1}{\sqrt{2}}\big(f_1'(\beta(x^k,t^k))+f_2'(\beta(x^k,t^k))\big)=(f\circ\beta)_t'(0,t).$$

For the derivative with respect to x_i we get

$$\lim_{k\to\infty}(f\circ\beta)_{x_i}'(x^k,t^k)=\lim_{k\to\infty}\frac{x_i^k}{\sqrt{2}\|x^k\|}\big(f_1'(\beta(x^k,t^k))-f_2'(\beta(x^k,t^k))\big)=0$$

because $x_i^k/\|x^k\|$ is bounded and the continuity of ∇f at $\beta(0,t)$ gives us

$$\lim_{k\to\infty}(f_1'(\beta(x^k,t^k))-f_2'(\beta(x^k,t^k)))=f_1'(\beta(0,t))-f_2'(\beta(0,t))=0.$$

The last equality follows from the fact that f is symmetric.

The opposite direction of the theorem is easy to prove by using (1.2). □

5. The decomposition functions. In this section we define the functions d_z and d_z^* and summarize some of their properties which will be used frequently. We call them *decomposition* functions because they will be used to describe how the subgradients of $f\circ\beta$ are decomposed into subgradients of f.

DEFINITION 5.1. *For every nonzero vector z in $\mathbb{R}^n$, we define the map*

$$d_z : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^2,$$
$$d_z(y,t) = \frac{1}{\sqrt{2}}\left(t + \frac{z^T y}{\|z\|}, t - \frac{z^T y}{\|z\|}\right).$$

In cases when the direction (y,t) is fixed and clear from the context we simply write d_z instead of $d_z(y,t)$.

DEFINITION 5.2. *For every nonzero vector z in $\mathbb{R}^n$, we define the map*

$$d_z^* : \mathbb{R}^2 \to \mathbb{R}^n \times \mathbb{R},$$
$$d_z^*(a,b) = \left(\frac{z}{\|z\|}\frac{a-b}{\sqrt{2}}, \frac{a+b}{\sqrt{2}}\right).$$

The following lemma collects a few elementary properties of the maps d_z and d_z^*. The proof is omitted.

LEMMA 5.3. *Let z and w be nonzero vectors in $\mathbb{R}^n$.*

(i) *The maps $d_z(\cdot)$ and $d_z^*(\cdot)$ are linear and adjoint to each other.*

(ii) *For every point (γ_1, γ_2) in $\mathbb{R}^2$,*

$$d_w d_z^*(\gamma_1,\gamma_2) = \frac{1+\delta}{2}(\gamma_1,\gamma_2) + \frac{1-\delta}{2}(\gamma_2,\gamma_1),$$

where $\delta = \frac{w^T z}{\|w\|\|z\|} \in [-1,1]$. In particular, when $w = z$ we have

$$d_z d_z^*(\gamma_1,\gamma_2) = (\gamma_1,\gamma_2).$$

(iii) *For every point (y,r) in $\mathbb{R}^n \times \mathbb{R}$ such that $y = az$ for some $a \in \mathbb{R}$,*

$$d_z^* d_z(y,r) = (y,r).$$

LEMMA 5.4. *Let A and B be symmetric subsets of $\mathbb{R}^2$. The sets*

$$\mathcal{D}(A) = \{d_z^*(\gamma_1,\gamma_2) | (\gamma_1,\gamma_2) \in A,\ z \neq 0\},$$
$$\mathcal{C}(A) = \{(y,r) | d_z(y,r) \in A \text{ for all } z \neq 0\}$$

satisfy the following properties.

(i) *If A is convex, then*
  (a) *if (x,t) is in $\mathcal{D}(A)$, then $(\delta x, t)$ is in $\mathcal{D}(A)$ for every $\delta \in [-1,1]$.*
  (b) *$\mathcal{D}(A)$ is a convex set.*
  (c) *$\mathcal{D}(A) = \mathcal{C}(A)$.*
  (d) *If B is also convex, then $\operatorname{cl}(\mathcal{D}(A) + \mathcal{D}(B)) = \operatorname{cl}\mathcal{D}(A+B)$.*

(ii) *For any A we have*
  (a) $\operatorname{conv}\mathcal{D}(A) = \mathcal{D}(\operatorname{conv} A)$.
  (b) $\mathcal{D}(\operatorname{cl} A) = \operatorname{cl}\mathcal{D}(A)$.

Proof. Part (i)(a). Let $(x,t) = d_z^*(\gamma_1,\gamma_2)$ for some (γ_1,γ_2) in A and $z \neq 0$. Since the set A is symmetric and convex, (γ_2,γ_1) is in A, and for every $\alpha \in [0,1]$ the convex combination $(\alpha\gamma_1 + (1-\alpha)\gamma_2, \alpha\gamma_2 + (1-\alpha)\gamma_1)$ is in A. Thus,

$$d_z^*\big(\alpha\gamma_1 + (1-\alpha)\gamma_2, \alpha\gamma_2 + (1-\alpha)\gamma_1\big) = \left(\frac{z}{\|z\|}\frac{\gamma_1-\gamma_2}{\sqrt{2}}(2\alpha-1), \frac{\gamma_1+\gamma_2}{\sqrt{2}}\right)$$
$$= (x(2\alpha-1), t) \in \mathcal{D}$$

for all $\alpha \in [0,1]$. Now set $\delta := 2\alpha - 1$.

Part (i)(b). Since A is convex, for any two points (γ_1, γ_2) and (δ_1, δ_2) in A and $\mu \in [0,1]$, we have that $(\mu\gamma_1 + (1-\mu)\delta_1, \mu\gamma_2 + (1-\mu)\delta_2)$ is in A. Thus, for every $z \neq 0$,

$$\left(\frac{z}{\|z\|} \frac{\mu(\gamma_1 - \gamma_2) + (1-\mu)(\delta_1 - \delta_2)}{\sqrt{2}}, \frac{\mu(\gamma_1 + \gamma_2) + (1-\mu)(\delta_1 + \delta_2)}{\sqrt{2}} \right) \in \mathcal{D}. \tag{5.1}$$

Take two points (x^1, t^1) and (x^2, t^2) in $\mathcal{D}$ and a number $\mu \in (0,1)$. We want to show that $(\mu x^1 + (1-\mu)x^2, \mu t^1 + (1-\mu)t^2)$ is also in $\mathcal{D}$. Suppose

$$(x^1, t^1) = d^*_{z^1}(\gamma_1, \gamma_2), \qquad (x^2, t^2) = d^*_{z^2}(\delta_1, \delta_2)$$

for some (γ_1, γ_2) and (δ_1, δ_2) in A, $z^1 \neq 0$, and $z^2 \neq 0$. Set

$$z_\mu := \mu \frac{\gamma_1 - \gamma_2}{\sqrt{2}} \frac{z^1}{\|z^1\|} + (1-\mu) \frac{\delta_1 - \delta_2}{\sqrt{2}} \frac{z^2}{\|z^2\|}$$

and notice that

$$\|z_\mu\| \leq \mu \frac{|\gamma_1 - \gamma_2|}{\sqrt{2}} + (1-\mu) \frac{|\delta_1 - \delta_2|}{\sqrt{2}}.$$

Then

$$\mu(x^1, t^1) + (1-\mu)(x^2, t^2) = \left(z_\mu, \frac{\mu(\gamma_1 + \gamma_2) + (1-\mu)(\delta_1 + \delta_2)}{\sqrt{2}} \right). \tag{5.2}$$

If $z_\mu = 0$, then from (5.1) and part (i)(a) with $\delta = 0$ we see that

$$\mu(x^1, t^1) + (1-\mu)(x^2, t^2) \in \mathcal{D}.$$

Suppose now that $z_\mu \neq 0$. Choose one of the points (γ_1, γ_2), (γ_2, γ_1) in A and one of the points (δ_1, δ_2), (δ_2, δ_1) in A so that, using part (i)(a), inclusion (5.1) becomes

$$\left(\frac{z}{\|z\|} \frac{\mu|\gamma_1 - \gamma_2| + (1-\mu)|\delta_1 - \delta_2|}{\sqrt{2}} \delta, \frac{\mu(\gamma_1 + \gamma_2) + (1-\mu)(\delta_1 + \delta_2)}{\sqrt{2}} \right) \in \mathcal{D}$$

for all $z \neq 0$ and $\delta \in (0,1)$. Let δ be a number in $(0,1)$ such that

$$\frac{\mu|\gamma_1 - \gamma_2| + (1-\mu)|\delta_1 - \delta_2|}{\sqrt{2}} \delta = \|z_\mu\|.$$

Putting this all together we obtain that (5.2) is in $\mathcal{D}$, showing that $\mathcal{D}$ is a convex set.

Part (i)(c). Suppose $(y, r) \in \mathcal{C}$; then $d_z(y, r) \in A$ for all $z \neq 0$. Apply Lemma 5.3(iii) with $a = 0$ and any z if $y = 0$, or with $a = 1$ and $z = y$ if $y \neq 0$, to obtain

$$(y, r) = d^*_z d_z(y, r) = d^*_z(d_z(y, r)) \in \mathcal{D}.$$

This shows that $\mathcal{C} \subseteq \mathcal{D}$.

Suppose now that $(y, r) \in \mathcal{D}$. That is, $(y, r) = d^*_z(\gamma_1, \gamma_2)$ for some (γ_1, γ_2) in A and some $z \neq 0$. Let $\hat{z}$ be an arbitrary nonzero vector and set $\delta := \frac{z^T \hat{z}}{\|z\|\|\hat{z}\|} \in [-1, 1]$. Then by Lemma 5.3(ii) we have

$$d_{\hat{z}}(y, r) = d_{\hat{z}} d^*_z(\gamma_1, \gamma_2) = \frac{1+\delta}{2}(\gamma_1, \gamma_2) + \frac{1-\delta}{2}(\gamma_2, \gamma_1) \in A$$

because A is symmetric and convex. Thus, $\mathcal{D} \subseteq \mathcal{C}$.

Part (i)(d). By part (i)(b) we have that both $\mathcal{D}(A) + \mathcal{D}(B)$ and $\mathcal{D}(A + B)$ are convex sets. It is clear that the latter set is contained in the former:

$$\operatorname{cl}(\mathcal{D}(A) + \mathcal{D}(B)) \supseteq \operatorname{cl}\mathcal{D}(A + B).$$

In order to show that the sets are equal it suffices to show that their support functions are equal. Fix any $x \in \mathbb{R}^n$ and suppose first that $x \neq 0$. In the first and last equality below, we use the fact that A and B are symmetric sets:

$$\begin{aligned}
&\max\{\langle (x,t), (d^*_{z^1}(\gamma_1,\gamma_2) + d^*_{z^2}(\delta_1,\delta_2))\rangle | (\gamma_1,\gamma_2) \in A,\ (\delta_1,\delta_2) \in B,\ z^1 \neq 0,\ z^2 \neq 0\} \\
&\quad = \max\{\langle (x,t), (d^*_x(\gamma_1,\gamma_2) + d^*_x(\delta_1,\delta_2))\rangle | (\gamma_1,\gamma_2) \in A,\ (\delta_1,\delta_2) \in B\} \\
&\quad = \max\{\langle (x,t), d^*_x(\gamma_1+\delta_1, \gamma_2+\delta_2)\rangle | (\gamma_1,\gamma_2) \in A,\ (\delta_1,\delta_2) \in B\} \\
&\quad = \max\{\langle (x,t), d^*_x(\alpha_1,\alpha_2)\rangle | (\alpha_1,\alpha_2) \in A + B\} \\
&\quad = \max\{\langle (x,t), d^*_z(\gamma_1,\gamma_2)\rangle | (\gamma_1,\gamma_2) \in A + B,\ z \neq 0\}.
\end{aligned}$$

The case $x = 0$ is easy.

Part (ii)(a). The inclusion $A \subseteq \operatorname{conv} A$ implies $\mathcal{D}(A) \subseteq \mathcal{D}(\operatorname{conv} A)$. Since the set $\mathcal{D}(\operatorname{conv} A)$ is convex by part (i)(b), we obtain $\operatorname{conv}\mathcal{D}(A) \subseteq \mathcal{D}(\operatorname{conv} A)$. The opposite inclusion $\mathcal{D}(\operatorname{conv} A) \subseteq \operatorname{conv}\mathcal{D}(A)$ is easy.

Part (ii)(b). Let $\{d^*_{x^k}(\gamma_1^k, \gamma_2^k)\}$ be a sequence in $\mathcal{D}(A)$ approaching a vector (z, s). Since the unit sphere in $\mathbb{R}^n$ is compact, we can find a subsequence, denoted again by k, such that $x^k/\|x^k\|$ converges to a unit vector x. For this subsequence we have $|\gamma_1^k - \gamma_2^k| \to \sqrt{2}\|z\|$ and $\gamma_1^k + \gamma_2^k \to \sqrt{2}s$. Consequently, $\{(\gamma_1^k, \gamma_2^k)\}$ is bounded so there is a subsequence, denoted again by k, for which $(\gamma_1^k, \gamma_2^k) \to (\gamma_1, \gamma_2) \in \operatorname{cl} A$. So, the sequence $\{d^*_{x^k}(\gamma_1^k, \gamma_2^k)\}$ approaches $d^*_x(\gamma_1, \gamma_2)$ which is in $\mathcal{D}(\operatorname{cl} A)$. This shows that for an arbitrary set A we have the inclusion $\mathcal{D}(\operatorname{cl} A) \supseteq \operatorname{cl}\mathcal{D}(A)$. The opposite inclusion is easy. □

6. Clarke subdifferential: The Lipschitz case. Suppose that h is a real valued function defined on some subset of $\mathbb{R}^m$. We say that h is *locally Lipschitz* at x in $\mathbb{R}^m$ if there exists a scalar K such that

$$|h(x'') - h(x')| \leq K\|x'' - x'\| \quad \text{for all } x'', x' \text{ close to } x.$$

For locally Lipschitz functions, the *Clarke directional derivative* [4] at the point x in the direction v is defined as

$$h^\circ(x; v) = \limsup_{y \to x;\, \lambda \downarrow 0} \frac{h(y + \lambda v) - h(y)}{\lambda}.$$

For y close to x and λ close to 0, the difference quotient in the definition of $h^\circ(x; v)$ is bounded above by $K|v|$. Thus, $h^\circ(x; v)$ is well defined and finite. We need the following formula for the Clarke directional derivative, which can be found in [4, p. 64]:

$$h^\circ(x; v) = \limsup_{y \to x} \{\langle \nabla h(y), v\rangle \mid y \text{ is such that } \nabla h(y) \text{ exists}\}^1 \tag{6.1}$$

for every pair $(x; v)$. In other words, there exists a sequence $\{x^k\}$ in $\mathbb{R}^m$ approaching x such that f is differentiable at each x_n and

$$\langle \nabla h(x^k), v\rangle \to h^\circ(x; v). \tag{6.2}$$

[1] By Rademacher's theorem, locally Lipschitz functions are differentiable almost everywhere.

The *Clarke subdifferential* $\partial^c h(x)$ is defined as

$$\partial^c h(x) = \{\xi \,|\, \langle v, \xi\rangle \le h^\circ(x; v) \text{ for all } v\}.$$

In can be shown that the set $\partial^c h(x)$ is compact, nonempty, and convex. If h is convex and finite on a neighborhood of x, then $\partial^c h(x) = \partial h(x)$, and if h is continuously differentiable at x, then $\partial^c h(x) = \{\nabla h(x)\}$. In this sense the Clarke subdifferential generalizes both the convex subdifferential and the gradient of a C^1 function. Finally, Proposition 2.1.2 in [4] shows that the Clarke directional derivative is the *support function* of the Clarke subdifferential:

$$h^\circ(x; v) = \max\{\langle v, \xi\rangle \,|\, \xi \in \partial^c h(x)\}. \tag{6.3}$$

Now, we return to the symmetric, bivariate function f, which we now require to be locally Lipschitz. It is not difficult to see that f is locally Lipschitz if and only if $f \circ \beta$ is locally Lipschitz. We will present a formula expressing the Clarke subdifferential of $f \circ \beta$ in terms of the Clarke subdifferential of f.

The following elementary lemma shows that the Clarke directional derivative of $h \circ \beta$ is invariant under Lorentz orthogonal transformations of the argument and the direction.

LEMMA 6.1. *Let (x, t) be a point in the domain of $f \circ \beta$, let (y, r) be a direction, and let U be an orthogonal matrix. Then*

$$(f \circ \beta)^\circ((x, t); (y, r)) = (f \circ \beta)^\circ((Ux, t); (Uy, r)).$$

THEOREM 6.2 (Clarke directional derivative). *Let $(0, t)$ be a point in the domain of $f \circ \beta$ and let (y, r) be any direction. Then if $x = 0$,*

$$(f \circ \beta)^\circ((0, t); (y, r)) = \max\{f^\circ(\beta(0, t); d_z(y, r)) | z \in \mathbb{R}^n,\ z \ne 0\}. \tag{6.4}$$

Note 6.3. For the Clarke directional derivative at a point (x, t) with $x \ne 0$, see Corollary 6.6.

Proof. By (6.2), there is a sequence of points $\{(x^k, t^k)\}$ approaching $(0, t)$ such that

$$(f \circ \beta)^\circ((0, t); (y, r)) = \lim_{k\to\infty} \langle \nabla(f \circ \beta)(x^k, t^k), (y, r)\rangle.$$

In order to evaluate $\nabla(f \circ \beta)$ using Theorem 4.1 we need to consider two cases, depending on whether x^k is zero or not. The general situation follows from these two cases by considering subsequences.

Case 1.a. Suppose $x^k = 0$ for all k. Let $\beta^k := \beta(0, t^k)$ and note that $f_1'(\beta^k) = f_2'(\beta^k)$. Fix an arbitrary nonzero vector $z \in \mathbb{R}^n$. Then

$$\begin{aligned}
(f \circ \beta)^\circ((0, t); (y, r)) &= \lim_{k\to\infty} \langle \nabla(f \circ \beta)(0, t^k), (y, r)\rangle \\
&= \lim_{k\to\infty} \left\langle \left(0, \dots, 0, \frac{f_1'(\beta^k) + f_2'(\beta^k)}{\sqrt{2}}\right), (y, r) \right\rangle \\
&= \lim_{k\to\infty} \langle \nabla f(\beta^k), \beta(0, r)\rangle \\
&= \lim_{k\to\infty} \left\langle \nabla f(\beta^k), d_z(y, r)\right\rangle \\
&\le f^\circ(\beta(0, t); d_z(y, r)).
\end{aligned}$$

In the last inequality we used (6.1).

Case 1.b. Suppose $x^k \neq 0$ for all k, $\lim_{k\to\infty} \frac{x^k}{\|x^k\|} = \frac{z}{\|z\|}$, and let $\beta^k := \beta(x^k, t^k)$. Then we have

$$\begin{aligned}
(f\circ\beta)^\circ((0,t);(y,r)) &= \lim_{k\to\infty} \langle \nabla(f\circ\beta)(x^k,t^k),(y,r)\rangle \\
&= \lim_{k\to\infty} \left\langle \left(\frac{f_1'(\beta^k) - f_2'(\beta^k)}{\sqrt{2}\|x^k\|} x^k, \frac{f_1'(\beta^k)+f_2'(\beta^k)}{\sqrt{2}} \right), (y,r) \right\rangle \\
&= \lim_{k\to\infty} \frac{f_1'(\beta^k) - f_2'(\beta^k)}{\sqrt{2}\|x^k\|} (x^k)^T y + \frac{f_1'(\beta^k)+f_2'(\beta^k)}{\sqrt{2}} r \\
&= \lim_{k\to\infty} f_1'(\beta^k)\left(\frac{r}{\sqrt{2}} + \frac{(x^k)^T y}{\sqrt{2}\|x^k\|} \right) + f_2'(\beta^k)\left(\frac{r}{\sqrt{2}} - \frac{(x^k)^T y}{\sqrt{2}\|x^k\|} \right) \\
&= \lim_{k\to\infty} \left\langle \nabla f'(\beta^k), d_z(y,r) \right\rangle \\
&\leq f^\circ(\beta(0,t); d_z(y,r)),
\end{aligned}$$

where, in substituting $x^k/\|x^k\|$ by $z/\|z\|$ in the last equality, we used the fact that since f is locally Lipschitz the sequence $\{(f_1'(\beta^k), f_2'(\beta^k)\}$ is bounded. All this shows that if $x = 0$, then

$$(f\circ\beta)^\circ((0,t);(y,r)) \leq \sup\{f^\circ(\beta(0,t); d_z(y,r)) | z \in \mathbb{R}^n,\ z \neq 0\}.$$

To show the opposite inequality, fix a nonzero vector $z \in \mathbb{R}^n$. There is a sequence of points $\{(a_k, b_k)\}$ approaching $\beta(0,t)$ such that

$$f^\circ(\beta(0,t); d_z(y,r)) = \lim_{n\to\infty} \langle \nabla f(a_k,b_k), d_z(y,r)\rangle.$$

There is an infinite subsequence $\{(a_{k'}, b_{k'})\}$ of $\{(a_k,b_k)\}$ that satisfies one of the following three possibilities:

(i) $a_{k'} = b_{k'}$ for all k'.
(ii) $a_{k'} > b_{k'}$ for all k'.
(iii) $a_{k'} < b_{k'}$ for all k'.

For this subsequence we still have

$$f^\circ(\beta(0,t); d_z(y,r)) = \lim_{k'\to\infty} \langle \nabla f(a_{k'}, b_{k'}), d_z(y,r)\rangle.$$

Without loss of generality, we may assume that $\{(a_k,b_k)\}$ satisfies one of the three possibilities and we may consider them separately.

Case 2.a. Suppose $a_k = b_k$ for all k. Note that in this case we have $f_1'(a_k,a_k) = f_2'(a_k,a_k)$. Thus,

$$\begin{aligned}
f^\circ(\beta(0,t); d_z(y,r)) &= \lim_{k\to\infty} \langle \nabla f(a_k,a_k), d_z(y,r)\rangle \\
&= \lim_{k\to\infty} \frac{f_1'(a_k,a_k)+f_2'(a_k,a_k)}{\sqrt{2}} r \\
&= \lim_{k\to\infty} \langle \nabla(f\circ\beta)(0,a_k),(y,r)\rangle \\
&\leq (f\circ\beta)^\circ((0,t);(y,r)).
\end{aligned}$$

Case 2.b. Suppose $a_k > b_k$ for all k. Define the sequence of vectors

$$z^k := \left(\frac{a_k - b_k}{2}, 0, \ldots, 0 \right) \in \mathbb{R}^n$$

(notice that $\|z^k\| = (a_k - b_k)/2$) and let U be an orthogonal matrix such that

$$\lim_{k\to\infty} \frac{Uz^k}{\|z^k\|} = \frac{z}{\|z\|}. \tag{6.5}$$

In the third equality below, we use the fact that the Lipschitzness of f implies that the sequence $\{f_1'(a_k,b_k) - f_2'(a_k,b_k)\}$ is bounded, and thus in the limit we can replace $z/\|z\|$ by $Uz^k/\|z^k\|$. We calculate

$$\begin{aligned}
f^\circ(\beta(0,t); d_z(y,r)) &= \lim_{k\to\infty} \langle \nabla f(a_k,b_k), d_z(y,r)\rangle \\
&= \lim_{k\to\infty} \left\langle \frac{f_1'(a_k,b_k) - f_2'(a_k,b_k)}{\sqrt{2}\|z\|} z, y \right\rangle + \frac{f_1'(a_k,b_k) + f_2'(a_k,b_k)}{\sqrt{2}} r \\
&= \lim_{k\to\infty} \left\langle \frac{f_1'(a_k,b_k) - f_2'(a_k,b_k)}{\sqrt{2}\|z^k\|} z^k, U^T y \right\rangle + \frac{f_1'(a_k,b_k) + f_2'(a_k,b_k)}{\sqrt{2}} r \\
&= \lim_{k\to\infty} \left\langle \nabla(f\circ\beta)\left(\frac{a_k - b_k}{\sqrt{2}}, 0, \ldots, 0, \frac{a_k + b_k}{\sqrt{2}}\right), (U^T y, r) \right\rangle \\
&\le (f\circ\beta)^\circ((0,t); (U^T y, r)) \\
&= (f\circ\beta)^\circ((0,t); (y,r)).
\end{aligned}$$

In the last equality we used Lemma 6.1.

Case 2.c. Suppose $a_k < b_k$ for all k. Define the sequence of vectors

$$z^k := \left(\frac{b_k - a_k}{2}, 0, \ldots, 0\right) \in \mathbb{R}^n$$

(notice that $\|z^k\| = (b_k - a_k)/2$) and proceed analogously to the previous case. ∎

It is straightforward to check that for every $(y,r) \in \mathbb{R}^n \times \mathbb{R}$, and every nonzero $x \in \mathbb{R}^n$, we have

$$\lim_{(x',t')\to(x,t),\, \mu\downarrow 0} \frac{\beta((x',t') + \mu(y,r)) - \beta(x',t')}{\mu} = d_x(y,r).$$

Applying [3, Theorem 6.2.3] to the Lipschitz map $\beta(x,t)$, we obtain the following result.

LEMMA 6.4. *If $x \neq 0$, then $\beta(x,t)$ is strictly differentiable and its strict derivative is the linear map d_x. That is,*

$$\lim_{\substack{(x',t'),\,(x'',t'')\to(x,t)\\ (x',t')\neq(x'',t'')}} \frac{\beta(x',t') - \beta(x'',t'') - d_x(x'-x'', t'-t'')}{\|(x'-x'', t'-t'')\|} = 0.$$

We now turn our attention to the problem of characterizing the Clarke subdifferential $\partial^c(f\circ\beta)(x,t)$.

THEOREM 6.5. *The Clarke subgradient at (x,t) of a Lorentz invariant function $f\circ\beta$, locally Lipschitz at (x,t), is given by the following formulae:*

(i) *If $x \neq 0$, then*

$$\partial^c(f\circ\beta)(x,t) = \{d_x^*(\gamma_1,\gamma_2) \,|\, (\gamma_1,\gamma_2) \in \partial^c f(\beta(x,t))\};$$

(ii) *if $x = 0$, then*

$$\partial^c(f\circ\beta)(0,t) = \{d_z^*(\gamma_1,\gamma_2) \,|\, (\gamma_1,\gamma_2) \in \partial^c f(\beta(0,t)),\ z \neq 0\}.$$

Proof. Case (i). Suppose that $x \neq 0$. Then, by Lemma 6.4, β is strictly differentiable at (x,t) with strict derivative d_x. Moreover, d_x is a surjective linear map. So we can apply the chain rule for the Clarke subdifferential [4, Theorem 2.3.10], which in our situation holds with equality:

$$\partial^c(f \circ \beta)(x,t) = \partial^c f(\beta(x,t)) \circ d_x.$$

Now, if $(v,p) \in \partial^c(f \circ \beta)(x,t)$ and $(y,r) \in \mathbb{R}^n \times \mathbb{R}$, then there is a subgradient $(\gamma_1,\gamma_2) \in \partial^c f(\beta(x,t))$ such that

$$\langle (v,p),(y,r)\rangle = ((\gamma_1,\gamma_2) \circ d_x)(y,r) = \langle (\gamma_1,\gamma_2), d_x(y,r)\rangle = \langle d_x^*(\gamma_1,\gamma_2),(y,r)\rangle,$$

where the last equality follows by Lemma 5.3. So

$$\partial^c(f \circ \beta)(x,t) \subseteq \{d_x^*(\gamma_1,\gamma_2) | (\gamma_1,\gamma_2) \in \partial^c f(\beta(x,t))\};$$

the other inclusion is now clear.

Case (ii). Suppose that $x = 0$ and define

$$\mathcal{D} := \{d_z^*(\gamma_1,\gamma_2) | (\gamma_1,\gamma_2) \in \partial^c f(\beta(0,t)),\ z \neq 0\}.$$

Two closed, convex sets are equal whenever their support functions are the same. The support function for the set $\operatorname{conv}\mathcal{D}$, evaluated at (y,r), is

$$\begin{aligned}
&\max\{\langle (y,r),(z,s)\rangle \,|\, (z,s) \in \operatorname{conv}\mathcal{D}\} \\
&\quad = \max\{\langle (y,r),(z,s)\rangle \,|\, (z,s) \in \mathcal{D}\} \\
&\quad = \max\{\langle (y,r), d_z^*(\gamma_1,\gamma_2)\rangle \,|\, (\gamma_1,\gamma_2) \in \partial^c f(\beta(0,t)),\ z \neq 0\} \\
&\quad = \max\{\langle d_z(y,r),(\gamma_1,\gamma_2)\rangle \,|\, (\gamma_1,\gamma_2) \in \partial^c f(\beta(0,t)),\ z \neq 0\} \\
&\quad = \max\{\max\{\langle d_z(y,r),(\gamma_1,\gamma_2)\rangle \,|\, (\gamma_1,\gamma_2) \in \partial^c f(\beta(0,t))\} \,|\, z \neq 0\} \\
&\quad = \max\{f^\circ(\beta(0,t); d_z(y,r)) \,|\, z \neq 0\} \\
&\quad = (f \circ \beta)^\circ((0,t);(y,r)),
\end{aligned}$$

where, in the last two equalities, we used (6.3) and Theorem 6.2. By (6.3), again applied to the function $f \circ \beta$, we obtain

$$\operatorname{cl}\operatorname{conv}\mathcal{D} = \partial^c(f \circ \beta)(0,t)$$

because $\partial^c(f \circ \beta)(x,t)$ is a closed convex set [4, Proposition 2.1.2]. The fact that $\operatorname{conv}\mathcal{D} = \mathcal{D}$ follows from Lemma 5.4(i)(b) and the fact that $\mathcal{D}$ is closed follows by Lemma 5.4(ii)(b). □

COROLLARY 6.6 (Clarke directional derivative). *Let (x,t) be a point in the domain of $f \circ \beta$ and let (y,r) be a direction in $\mathbb{R}^n \times \mathbb{R}$. Then if $x \neq 0$,*

$$(f \circ \beta)^\circ((x,t);(y,r)) = f^\circ(\beta(x,t); d_x(y,r)).$$

Proof. Use again the fact that $(f \circ \beta)^\circ((x,t);(y,r))$ is the support function of $\partial^c(f \circ \beta)(x,t)$; see [4, Proposition 2.1.2]. □

7. Second-order properties. In this section, we let f be twice differentiable at the point (a,b). This means that f is differentiable in a neighborhood of this point, and the first derivative, ∇f, is differentiable again at (a,b). The question that we are going to answer now is whether $g := f \circ \beta$ is twice differentiable at any point (x,t) such that $\beta(x,t) = (a,b)$. Elementary calculus shows that, if $x \neq 0$, then g is twice differentiable. It turns out that this is always the case, as we prove in Theorem 7.1. A generalization of Theorems 7.1 and 7.2 to the setting of formally real Jordan algebras can be found in [25]. Our approach is direct and first appeared in [21].

7.1. Second-order differentiability.

THEOREM 7.1. *The function $g := f \circ \beta$ is twice differentiable at (x,t) if and only if f is twice differentiable at $\beta(x,t)$. In that case, we have the following:*

(i) *If $x \neq 0$, then*

$$\begin{aligned}
g''_{x_i x_j}(x,t) &= \frac{x_i x_j}{2\|x\|^2}(f''_{11} - f''_{12} - f''_{21} + f''_{22}) + \frac{\delta_{ij}\|x\|^2 - x_i x_j}{\sqrt{2}\|x\|^3}(f'_1 - f'_2),\\
g''_{tx_i}(x,t) &= \frac{x_i}{2\|x\|}(f''_{11} - f''_{12} + f''_{21} - f''_{22}),\\
g''_{x_i t}(x,t) &= \frac{x_i}{2\|x\|}(f''_{11} + f''_{12} - f''_{21} - f''_{22}),\\
g''_{tt}(x,t) &= \frac{1}{2}(f''_{11} + f''_{12} + f''_{21} + f''_{22}),
\end{aligned}$$

where δ_{ij} is 1 if $i = j$ and 0 otherwise;

(ii) *if $x = 0$, then*

$$\begin{aligned}
g''_{x_i x_j}(0,t) &= \begin{cases} \frac{1}{2}(f''_{11} - f''_{12} - f''_{21} + f''_{22}) & \text{if } i = j,\\ 0 & \text{otherwise,}\end{cases}\\
g''_{tx_i}(0,t) &= 0,\\
g''_{x_i t}(0,t) &= 0,\\
g''_{tt}(0,t) &= \frac{1}{2}(f''_{11} + f''_{12} + f''_{21} + f''_{22}).
\end{aligned}$$

All second-order derivatives of f, in both cases, are evaluated at $\beta(x,t)$.

Proof. The "only if" part follows easily from (1.2).

The verification of part (i) is straightforward. For part (ii) denote

$$\begin{aligned}
H_{ii} &:= \frac{1}{2}(f''_{11} - f''_{12} - f''_{21} + f''_{22}) \ \text{ for } i = 1,\dots,n,\\
H_{tt} &:= \frac{1}{2}(f''_{11} + f''_{12} + f''_{21} + f''_{22}),\\
H &:= \operatorname{Diag}(H_{11},\dots,H_{nn},H_{tt}),
\end{aligned}$$

where the second-order derivatives of f are evaluated at $\beta(0,t)$ and the operator Diag forms a diagonal matrix from its vector argument. Fix an arbitrary sequence $\{h^k\}$ in $\mathbb{R}^n \times \mathbb{R}$ converging to 0 and denote $\bar{h}^k := (h_1^k,\dots,h_n^k)^T$. Using Theorem 4.1 we show that the limit of the difference quotient

$$\lim_{k\to\infty} \frac{\|\nabla g(\bar{h}^k, t + h_{n+1}^k) - \nabla g(0,t) - Hh^k\|}{\|h^k\|}$$

is 0. We consider separately each coordinate in the difference quotient. Two cases are necessary: one for the coordinates from 1 to n and one for the $(n+1)$st coordinate. The sequence $\{h^k\}$ can be partitioned into two subsequences—one in which $\bar{h}^k = 0$ for all k and one in which $\bar{h}^k \neq 0$ for all k. We are done if we show that the limit of the difference quotient for each of the two subsequences is zero. That leads us to consider the following two subcases in each main case.

Subcase (a). Suppose $i \in \{1,\dots,n\}$. Then the difference quotient becomes

$$\lim_{k\to\infty} \frac{|g'_i(\bar{h}^k, t + h_{n+1}^k) - g'_i(0,t) - H_{ii}h_i^k|}{\|h^k\|}. \tag{7.1}$$

We use Theorem 4.1 to evaluate the derivatives g'_i. Notice that if $\bar{h}^k = 0$ for all k, then the limit is clearly 0. Thus, suppose $\bar{h}^k \neq 0$ for all k. Then (7.1) becomes

$$\lim_{k\to\infty} \frac{|\frac{\bar{h}^k}{\sqrt{2}\|\bar{h}^k\|}(f'_1(\beta(\bar{h}^k, t+h^k_{n+1})) - f'_2(\beta(\bar{h}^k, t+h^k_{n+1}))) - \frac{\bar{h}^k}{2}(f''_{11} - f''_{12} - f''_{21} + f''_{22})|}{\|h^k\|},$$

where the second derivatives of f are evaluated at $\beta(0,t)$. Because f'_1 and f'_2 exist in a neighborhood of $\beta(0,t)$ and are differentiable at $\beta(0,t)$, we have

$$f'_1(\beta(\bar{h}^k, t+h^k_{n+1})) \sim f'_1(\beta(0,t)) + f''_{11}(\beta(0,t))\frac{h^k_{n+1} + \|\bar{h}^k\|}{\sqrt{2}} + f''_{12}(\beta(0,t))\frac{h^k_{n+1} - \|\bar{h}^k\|}{\sqrt{2}},$$
$$f'_2(\beta(\bar{h}^k, t+h^k_{n+1})) \sim f'_2(\beta(0,t)) + f''_{21}(\beta(0,t))\frac{h^k_{n+1} + \|\bar{h}^k\|}{\sqrt{2}} + f''_{22}(\beta(0,t))\frac{h^k_{n+1} - \|\bar{h}^k\|}{\sqrt{2}},$$

where $\sim$ indicates that the difference of both sides is of order $o(\|h^k\|)$. Because f is symmetric, at the point $\beta(0,t)$ we have $f'_1 = f'_2$, $f''_{12} = f''_{21}$, and $f''_{11} = f''_{22}$. Substituting the two expansions into the limit shows that it is indeed 0.

Subcase (b). Suppose $i = n+1$. The arguments are analogous to the previous case. We use again Theorem 4.1 to evaluate the derivative g'_t and then substitute f'_1 and f'_2 with their first-order expansions. □

7.2. Continuity of the Hessian.

THEOREM 7.2. *The function $g := f \circ \beta$ is twice continuously differentiable at (x,t) if and only if f is at $\beta(x,t)$.*

Proof. The "only if" direction is also easy to obtain from (1.2). The "if" direction is clear in the case when $x \neq 0$. Thus, we suppose that f is twice continuously differentiable at $\beta(0,t)$ and we show that for any sequence of vectors $\{(x^k, t^k)\}$ in $\mathbb{R}^n \times \mathbb{R}$ approaching $(0,t)$, the Hessian $\nabla^2 g(x^k, t^k)$ is approaching $\nabla^2 g(0,t)$. Viewing, for a fixed basis, $\nabla^2 g(0,t)$ as a matrix, we prove the convergence for each entry. We again consider two cases, and the general situation follows easily from them. If $x^k = 0$ for all k, then the result follows directly from the continuity of $\nabla^2 f$ at the point $\beta(0,t)$; see Theorem 7.1. If $x^k \neq 0$ for all n, then from the continuity of $\nabla^2 f$ at the point $\beta(0,t)$ and the formulae given in Theorem 7.1, we have

$$\lim_{k\to\infty} g''_{x_i t}(x^k, t^k) = \lim_{k\to\infty} g''_{t x_i}(x^k, t^k) = 0,$$
$$\lim_{k\to\infty} g''_{tt}(x^k, t^k) = g''_{tt}(0,t),$$

where we also used the fact that since f is symmetric, $f''_{12} = f''_{21}$ and $f''_{11} = f''_{22}$ at the point $\beta(0,t)$. The interesting part is to show $\lim_{k\to\infty} g''_{x_i x_j}(x^k, t^k) = g''_{x_i x_j}(0,t)$. Denote

$$\beta^k_{+-} := \frac{1}{\sqrt{2}}(t^k + \|x^k\|, t^k - \|x^k\|),$$
$$\beta^k_{++} := \frac{1}{\sqrt{2}}(t^k + \|x^k\|, t^k + \|x^k\|),$$
$$\beta^k_{-+} := \frac{1}{\sqrt{2}}(t^k - \|x^k\|, t^k + \|x^k\|).$$

Because f is symmetric, $f_1'(\beta^k_{-+}) = f_2'(\beta^k_{+-})$. This allows us to evaluate the following limit using the mean value theorem:

$$\begin{aligned}\lim_{k\to\infty} \frac{1}{\sqrt{2}\|x^k\|}\big(f_1'(\beta(x^k,t^k)) - f_2'(\beta(x^k,t^k))\big)&\\ = \lim_{k\to\infty} \frac{1}{\sqrt{2}\|x^k\|}\big(f_1'(\beta^k_{+-}) - f_1'(\beta^k_{++}) + f_1'(\beta^k_{++}) - f_1'(\beta^k_{-+})\big)&\\ = \lim_{k\to\infty}\left(- f_{12}''\left(\frac{t^k+\|x^k\|}{\sqrt{2}},\nu^k\right) + f_{11}''\left(\mu^k,\frac{t^k+\|x^k\|}{\sqrt{2}}\right)\right)&\\ = \frac{1}{2}\big(f_{11}''(\beta(0,t)) - f_{12}''(\beta(0,t)) - f_{21}''(\beta(0,t)) + f_{22}''(\beta(0,t))\big).&\end{aligned}$$

Above, the numbers ν^k and μ^k are between $\frac{t^k-\|x^k\|}{\sqrt{2}}$ and $\frac{t^k+\|x^k\|}{\sqrt{2}}$, and the last equality uses the continuity of $\nabla^2 f$ and the fact that f is symmetric. Using the formula for $g''_{x_ix_j}$ given in Theorem 7.1, we can see that

$$\begin{aligned}\lim_{k\to\infty} g''_{x_ix_j}(x^k,t^k) &= \frac{\delta_{ij}}{2}\big(f_{11}''(\beta(0,t)) - f_{12}''(\beta(0,t)) - f_{21}''(\beta(0,t)) + f_{22}''(\beta(0,t))\big)\\ &= g''_{x_ix_j}(0,t).\end{aligned}$$

This concludes the proof. ∎

7.3. Positive definite Hessian. We begin with a simple lemma, and the main result of this subsection follows after it.

LEMMA 7.3. *Suppose that function f is continuously differentiable on an open convex subset of $\mathbb{R}^2$ and is strictly convex there. For any point (a,b) in its domain with $a > b$ we have $f_1'(a,b) > f_2'(a,b)$.*

THEOREM 7.4. *Suppose that f is twice continuously differentiable at $\beta(x,t)$. Then $\nabla^2(f\circ\beta)$ is positive definite at the point (x,t) if and only if $\nabla^2 f$ is positive definite at $\beta(x,t)$.*

Proof. Suppose that $\nabla^2 f(\beta(x,t))$ is positive definite. We use the formulae in Theorem 7.1 to give a matrix representation of the Hessian of $f\circ\beta$. Define the $(n+1)\times 2$ matrix:

$$X := \frac{1}{\sqrt{2}}\begin{pmatrix} \frac{x}{\|x\|} & -\frac{x}{\|x\|} \\ 1 & 1 \end{pmatrix}$$

and the $(n+1)\times(n+1)$ matrix:

$$M := \frac{1}{\sqrt{2}\|x\|}\begin{pmatrix} I_n - \frac{xx^T}{\|x\|^2} & 0 \\ 0 & 0 \end{pmatrix},$$

where I_n is the $n\times n$ identity matrix.

Case I. When $x \neq 0$, the Hessian of $f\circ\beta$ can be written as

$$\nabla^2(f\circ\beta)(x,t) = X\nabla^2 f(\beta(x,t))X^T + M\nabla f(\beta(x,t))\begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

For any nonzero vector (y,r), we have

$$\begin{aligned}(y,r)\big(\nabla^2(f\circ\beta)(x,t)\big)(y,r)^T &= \frac{1}{2}d_x(y,r)\nabla^2 f(\beta(x,t))d_x(y,r)^T\\ &\quad + \frac{1}{\sqrt{2}\|x\|^3}\big(\|y\|^2\|x\|^2 - (x^Ty)^2\big)\big(f_1'(\beta(x,t)) - f_2'(\beta(x,t))\big).\end{aligned}$$

Using Lemma 7.3 we see that the above expression is strictly positive.

Case II. In the case when $x = 0$, the Hessian of $f \circ \beta$ is a diagonal matrix, and the fact that it is positive definite can be easily seen.

In the other direction the result follows from (1.2). □

The proof of the next corollary is virtually the same as [22, Theorem 7.2], so we omit it.

COROLLARY 7.5. *Let C be a symmetric and convex subset of $\mathbb{R}^2$. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be a twice continuously differentiable function defined on C. Then*

$$\min_{(a,b)\in C} \lambda_{\min}(\nabla^2 f(a,b)) = \min_{(x,t)\in\beta^{-1}(C)} \lambda_{\min}(\nabla^2(f\circ\beta)(x,t)). \tag{7.2}$$

Above, $\lambda_{\min}$ denotes the smallest eigenvalue of the matrix in its argument.

Multiplying both sides of (7.2) by -1, we obtain

$$\max_{(a,b)\in C} \lambda_{\max}(\nabla^2 f(a,b)) = \max_{(x,t)\in\beta^{-1}(C)} \lambda_{\max}(\nabla^2(f\circ\beta)(x,t)).$$

8. The regular and proximal subdifferentials. Given a function $h : \mathbb{R}^m \to [-\infty, +\infty]$ and a point x in $\mathbb{R}^m$ at which h is finite, we call a vector y of $\mathbb{R}^m$ a *regular subgradient* of h at x if

$$h(x+z) \geq h(x) + \langle y, z\rangle + o(\|z\|) \text{ as } z \to 0.$$

The set of regular subgradients is denoted $\hat{\partial}h(x)$ and is called the *regular subdifferential* of h at x. If h is infinite at x, then the set $\hat{\partial}h(x)$ is defined to be empty. It is not difficult to show that it is always a closed and convex set; see [20].

A vector y is called a *proximal subgradient* of the function h at x, a point where $h(x)$ is finite, if there exist $\rho > 0$ and $\delta > 0$ such that

$$h(x+z) \geq h(x) + \langle y, z\rangle - \frac{1}{2}\rho\|z\|^2 \quad \text{when } \|z\| \leq \delta.$$

The set of all proximal subgradients will be denoted $\partial_p h(x)$. If h is infinite at x, then the set $\partial_p h(x)$ is defined to be empty. It is not difficult to show that it is always a closed and convex set; see [5].

Now let f be the symmetric, bivariate function on $\mathbb{R}^2$ and $g := f \circ \beta$. We are going to derive a formula for $\hat{\partial}g(x,t)$ in terms of $\hat{\partial}f(\beta(x,t))$. The next lemma lists several properties of the map $\beta(x,t)$ that we need. By $\mathbb{R}^n_{\geq}$ we denote the cone of vectors x in $\mathbb{R}^n$ satisfying $x_1 \geq x_2 \geq \cdots \geq x_n$.

LEMMA 8.1.

(i) *For any vector w in $\mathbb{R}^2_{\geq}$, the function $w^T\beta$ is convex and any point (x,t) in $\mathbb{R}^n \times \mathbb{R}$ satisfies $d^*_x(w) \in \partial(w^T\beta)(x,t)$.*

(ii) *The directional derivative $\beta'((x,t);(y,r))$ is given by*

$$\beta'((x,t);(y,r)) = \begin{cases} d_x(y,r) & \text{if } x \neq 0, \\ \beta(y,r) & \text{if } x = 0. \end{cases}$$

(iii) *The map β is Lipschitz with global constant* 1.

(iv) *Given a point (x,t) in $\mathbb{R}^n \times \mathbb{R}$, all vectors (z,s) close to zero satisfy*

$$\beta((x,t)+(z,s)) = \beta(x,t) + \beta'((x,t);(z,s)) + O(\|(z,s)\|^2).$$

Proof. (i) The convexity is elementary. To check the second half we need to verify that

$$w^T\beta(y,r) - w^T\beta(x,t) \geq \langle d_x^*(w_1,w_2), (y-x, r-t)\rangle,$$

which expanded and simplified is equivalent to

$$\frac{w_1-w_2}{\sqrt{2}}(\|y\| - \|x\|) \geq \frac{x^T(y-x)}{\|x\|}\frac{w_1-w_2}{\sqrt{2}}.$$

After cancellation, the last inequality follows from the Cauchy–Schwarz inequality.

(ii) This part is a straightforward verification.

(iii) For any points (x,t) and (z,s) we have

$$\begin{aligned}
&\|\beta((x,t)+(z,s)) - \beta(x,t)\| \\
&\quad = \frac{1}{\sqrt{2}}\|(t+s+\|x+z\|, t+s-\|x+z\|) - (t+\|x\|, t-\|x\|)\| \\
&\quad = \frac{1}{\sqrt{2}}\|(s+\|x+z\|-\|x\|, s-(\|x+z\|-\|x\|))\| \\
&\quad = \sqrt{s^2+(\|x+z\|-\|x\|)^2} \\
&\quad \leq \sqrt{s^2+\|z\|^2} \\
&\quad = \|(z,s)\|.
\end{aligned}$$

(iv) Suppose first that $x \neq 0$. Then using part (ii) of this lemma and using the Cauchy–Schwarz inequality several times we get

$$\begin{aligned}
&\|\beta((x,t)+(z,s)) - \beta(x,t) - \beta'((x,t);(z,s))\|^2 \\
&\qquad = \frac{1}{2}\left\|\left(\|x+z\| - \|x\| - \frac{x^Tz}{\|x\|}, -\|x+z\| + \|x\| + \frac{x^Tz}{\|x\|}\right)\right\|^2 \\
&\qquad = \left(\|x+z\| - \|x\| - \frac{x^Tz}{\|x\|}\right)^2 \\
&\qquad = O(\|z\|^4) = O(\|(z,s)\|^4),
\end{aligned}$$

where the penultimate equality holds since $\nabla\|\cdot\|(x) = \frac{x}{\|x\|}$.

The case $x = 0$ is easy. ∎

Let L be a subset of $\mathbb{R}^m$ and fix a point x in $\mathbb{R}^m$. An element d belongs to the *contingent cone* to L at x, denoted $K(L|x)$, if either $d = 0$ or there is a sequence $\{x^k\}$ in L approaching x with $(x^k - x)/\|x^k - x\|$ approaching $d/\|d\|$. The *negative polar* of a subset H of $\mathbb{R}^m$ is the set

$$H^- = \{y \in \mathbb{R}^m | \langle x, y\rangle \leq 0 \text{ for all } x \in H\}.$$

We use the following lemmas from [11]; see Propositions 2.1 and 2.2 there.

LEMMA 8.2. *Given a function $f : \mathbb{R}^m \to [-\infty, +\infty]$ and a point x^0 in $\mathbb{R}^m$, any regular subgradient of f at x^0 is polar to the contingent cone of the level set $L = \{x \in E : f(x) \leq f(x^0)\}$ at x^0; that is,*

$$\hat{\partial} f(x^0) \subset (K(L|x^0))^-.$$

LEMMA 8.3. *If the function $f : \mathbb{R}^m \to [-\infty, +\infty]$ is invariant under a subgroup G of $O(m)$, then any point x in $\mathbb{R}^m$ and transformation g in G satisfy $\hat{\partial} f(gx) = g\hat{\partial} f(x)$. Corresponding results hold for the proximal, approximate horizon and Clarke subgradients (see the next sections).*

We define the action of the orthogonal group $O(n)$ on $\mathbb{R}^n \times \mathbb{R}$ by

$$U.(x,t) = (Ux,t) \text{ for every } U \in O(n).$$

For a fixed point (x,t) in $\mathbb{R}^n \times \mathbb{R}$ we define the orbit

$$O(n).(x,t) = \{(Ux,t) | U \in O(n)\}.$$

If $x \neq 0$, this orbit is just an $n-1$ dimensional sphere with radius $\|x\|$ at level t in $\mathbb{R}^n \times \mathbb{R}$. So it is an $n-1$ dimensional manifold, and one can easily calculate that its tangent and normal spaces at the point (x,t) are

$$\begin{aligned} T_{(x,t)}(O(n).(x,t)) &= \{(y,0) | y^T x = 0\}, \\ N_{(x,t)}(O(n).(x,t)) &= \{(ax,b) | (a,b) \in \mathbb{R}^2\}. \end{aligned}$$

If $x = 0$, then

$$\begin{aligned} T_{(0,t)}(O(n).(0,t)) &= \{0\}, \\ N_{(0,t)}(O(n).(0,t)) &= \mathbb{R}^{n+1}. \end{aligned}$$

Now, using these observations and Lemma 8.2 we can say more about $\hat{\partial}(f \circ \beta)(x,t)$ in the case when $x \neq 0$.

LEMMA 8.4. *If $x \neq 0$ and $(y,r) \in \hat{\partial}(f \circ \beta)(x,t)$, then $(y,r) = (ax,r)$ for some $a \in \mathbb{R}$.*

Proof. If $(y,r) \in \hat{\partial}(f \circ \beta)(x,t)$, then by Lemma 8.2 we have

$$\begin{aligned} (y,r) &\in (K(\{(z,s) | (f \circ \beta)(z,s) \leq (f \circ \beta)(x,t)\} | (x,t)))^- \\ &\subset (K(O(n).(x,t) | (x,t)))^- \\ &= N_{(x,t)}(O(n).(x,t)). \end{aligned}$$

The claim follows from the expression for the normal space above. □

The following is the main theorem of this section.

THEOREM 8.5. *The regular subdifferential of any Lorentz invariant function $f \circ \beta$ at the point (x,t) is given by the following formulae:*

(i) *If $x \neq 0$, then*

$$\hat{\partial}(f \circ \beta)(x,t) = \{d_x^*(\gamma_1, \gamma_2) | (\gamma_1, \gamma_2) \in \hat{\partial} f(\beta(x,t))\};$$

(ii) *if $x = 0$, then*

$$\hat{\partial}(f \circ \beta)(0,t) = \{d_z^*(\gamma_1, \gamma_2) | (\gamma_1, \gamma_2) \in \hat{\partial} f(\beta(0,t)),\ z \neq 0\}.$$

Similar formulae hold for the proximal subdifferential.

Proof. Case (i). This case follows immediately from the chain rule [20, Exercise 10.7].

Case (ii). Let $x = 0$. We show that

$$\hat{\partial}(f \circ \beta)(0,t) = \{(y,r) | d_z(y,r) \in \hat{\partial} f(\beta(0,t)) \text{ for all } z \neq 0\}.$$

The stated version follows from Lemma 5.4(i)(c).

Suppose $(y,r) \in \hat{\partial}(f\circ\beta)(0,t)$, let $z := (z_1,z_2) \in \mathbb{R}^2$ be small, and let w be an arbitrary nonzero vector in $\mathbb{R}^n$. Then

$$\begin{aligned} f(\beta(0,t)+(z_1,z_2)) &= (f\circ\beta)\left((0,t)+\left(\frac{w}{\|w\|}\frac{z_1-z_2}{\sqrt{2}},\frac{z_1+z_2}{\sqrt{2}}\right)\right)\\ &\geq (f\circ\beta)(0,t)+\frac{w^Ty}{\|w\|}\frac{z_1-z_2}{\sqrt{2}}+r\frac{z_1+z_2}{\sqrt{2}}+o(\|z\|)\\ &= f(\beta(0,t))+\langle d_w(y,r),(z_1,z_2)\rangle+o(\|z\|). \end{aligned}$$

Consequently $d_w(y,r) \in \hat{\partial}f(\beta(0,t))$ for all $w \neq 0$.

In the opposite direction suppose that $d_w(y,r) \in \hat{\partial}f(\beta(0,t))$ for all $w \neq 0$. If $y = 0$, then for any vector $(z,s) \in \mathbb{R}^n \times \mathbb{R}$ close to 0 we have

$$\begin{aligned} (f\circ\beta)((0,t)+(z,s)) &= f(\beta(0,t)+(\beta((0,t)+(z,s))-\beta(0,t)))\\ &\geq f(\beta(0,t))+\langle d_w(0,r),(\beta((0,t)+(z,s))-\beta(0,t))\rangle+o(\|(z,s)\|)\\ &= f(\beta(0,t))+rs+o(\|(z,s)\|)\\ &= (f\circ\beta)(0,t)+\langle(0,r),(z,s)\rangle+o(\|(z,s)\|). \end{aligned}$$

Thus, $(0,r) \in \hat{\partial}(f\circ\beta)(0,t)$.

If $y \neq 0$, then for $w = y$ we have $d_y(y,r) \in \hat{\partial}f(\beta(0,t))$. Let $(z,s) \in \mathbb{R}^n \times \mathbb{R}$ be a vector close to 0. Then

$$\begin{aligned} (f\circ\beta)((0,t)+(z,s)) &= f(\beta(0,t)+(\beta((0,t)+(z,s))-\beta(0,t)))\\ &\geq f(\beta(0,t))+\langle d_y(y,r),(\beta((0,t)+(z,s))-\beta(0,t))\rangle+o(\|(z,s)\|)\\ &= f(\beta(0,t))+\|y\|\|z\|+rs+o(\|(z,s)\|)\\ &\geq (f\circ\beta)(0,t)+\langle(y,r),(z,s)\rangle+o(\|(z,s)\|). \end{aligned}$$

Consequently $(y,r) \in \hat{\partial}(f\circ\beta)(0,t)$.

The proof for the proximal subdifferential is essentially identical. □

9. The approximate and horizon subdifferentials. Given a function $h : \mathbb{R}^m \to [-\infty,+\infty]$ and a point x in $\mathbb{R}^m$ at which h is finite, a vector y of $\mathbb{R}^m$ is called an *approximate subgradient* of h at x if there is a sequence of points $\{x^k\}$ in $\mathbb{R}^m$ approaching x with values $h(x^k)$ approaching $h(x)$ and a sequence of regular subgradients y^k in $\hat{\partial}h(x^k)$ approaching y. The set of all approximate subgradients is called the *approximate subdifferential* $\partial h(x)$. A vector y is called a *horizon subgradient* if either $y = 0$ or there is a sequence of points $\{xv\}$ in $\mathbb{R}^m$ approaching x with values $h(x^k)$ approaching $h(x)$, a sequence $\{t^k\}$ of reals decreasing to zero, and a sequence of regular subgradients y^k in $\hat{\partial}h(x^k)$ for which t^ky^k approaches y. The set of all horizon subgradients is denoted $\partial^\infty h(x)$. If h is infinite at x, then the set $\partial h(x)$ is defined to be empty and $\partial^\infty h(x)$ to be $\{0\}$.

Recall that we used the same notation, $\partial h(x)$, for the convex subgradient when h is a convex function. There is no danger of confusion because the subdifferentials coincide when h is a proper, convex function; see [20, Proposition 8.12].

THEOREM 9.1. *The approximate subdifferential of any Lorentz invariant function $f\circ\beta$ at the point (x,t) is given by the following formulae:*

(i) *If $x \neq 0$, then*

$$\partial(f\circ\beta)(x,t) = \{d_x^*(a,b) \mid (a,b) \in \partial f(\beta(x,t))\};$$

(ii) *if* $x = 0$, *then*

$$\partial(f \circ \beta)(0,t) = \{d_z^*(a,b) \,|\, (a,b) \in \partial f(\beta(0,t)),\ z \neq 0\}.$$

Similar formulae hold for the horizon subgradient.

Proof. Part (i) $x \neq 0$. This case follows immediately from the chain rule [20, Exercise 10.7].

Part (ii) $x = 0$. Suppose $(y,r) \in \partial(f \circ \beta)(0,t)$. By definition, there is a sequence of points $\{(x^k,t^k)\}$ approaching $(0,t)$ with $(f\circ\beta)(x^k,t^k)$ approaching $(f\circ\beta)(0,t)$ and a sequence of regular subgradients $(y^k,r^k) \in \hat\partial(f\circ\beta)(x^k,t^k)$ approaching (y,r).

Case 1.a. Suppose $x^k = 0$ for all k. Then Theorem 8.5 says that $(y^k,r^k) = d^*_{z^k}(a_k,a_k)$ such that $(a_k,a_k) \in \hat\partial f(\beta(0,t^k))$ for some $z^k \neq 0$. Since (y^k,r^k) approaches (y,r) we get that $y = 0$ and $a_k \to a := r/\sqrt{2}$. Thus, $(0,r) = (0,\sqrt{2}a) = d_z^*(a,a)$ for any $z \neq 0$ and $(a,a) \in \partial f(\beta(0,t))$.

Case 1.b. Suppose $x^k \neq 0$ for all k. Then Theorem 8.5 says that $(y^k,r^k) = d^*_{x_q}(a_k,b_k)$ such that $(a_k,b_k) \in \hat\partial f(\beta(x^k,t^k))$. Let us choose a subsequence k' for which $x^{k'}/\|x^{k'}\|$ converges to a unit vector z. Then we have that $|a_{k'} - b_{k'}|$ approaches $\sqrt{2}\|y\|$ and $a_{k'} + b_{k'}$ approaches $\sqrt{2}r$, that is, $(a_{k'},b_{k'})$ is a bounded sequence, so if necessary we may choose a convergent subsequence k''. Then $(a_{k''},b_{k''}) \to (a,b) \in \partial f(\beta(0,t))$ and $(y,r) = d_z^*(a,b)$.

Case 1.c. Suppose the sequence x^k has infinitely many elements that are equal to 0 and infinitely many elements that are not equal to 0. Let $\{x^k\} = \{x^{k'}\} \cup \{x^{k''}\}$, where $x^{k'} \neq 0$ and $x^{k''} = 0$. We now choose any of the subsequences k' or k'' and apply the corresponding subcase above.

To show the opposite inclusion, suppose that $(y,r) = d_z^*(a,b)$ for some $(a,b) \in \partial f(\beta(0,t))$ and some $z \neq 0$. By the definition of approximate subgradients there is a sequence (c_k,d_k) approaching $\beta(0,t)$, with $f(c_k,d_k)$ approaching $f(\beta(0,t))$, and a sequence of regular subgradients (a_k,b_k) approaching (a,b) such that $(a_k,b_k) \in \hat\partial f(c_k,d_k)$. We have three possible cases.

Case 2.a. Suppose first that there is an infinite subsequence k' such that $c_{k'} > d_{k'}$ for all k'. Then $d_z^*(c_{k'},d_{k'})$ approaches $d_z^*(\beta(0,t)) = (0,t)$ with $f(c_{k'},d_{k'}) = (f\circ\beta)(d_z^*(c_{k'},d_{k'}))$ approaching $f(\beta(0,t)) = (f\circ\beta)(0,t)$ and regular subgradients $(a_{k'},b_{k'}) \in \hat\partial f(\beta(d_z^*(c_{k'},d_{k'})))$. If we set $z^{k'} := \frac{z}{\|z\|}\frac{c_{k'}-d_{k'}}{\sqrt{2}}$, then Theorem 8.5 says that $d^*_{z^{k'}}(a_{k'},b_{k'}) \in \hat\partial(f\circ\beta)(d_z^*(c_{k'},d_{k'}))$. Notice that $z^{k'}/\|z^{k'}\|$ converges to $z/\|z\|$, so $d^*_{z^{k'}}(a_{k'},b_{k'})$ approaches $d_z^*(a,b) = (y,r)$, and thus (y,r) is in $\partial(f\circ\beta)(0,t)$.

Case 2.b. There is an infinite subsequence k' such that $c_{k'} < d_{k'}$ for all k'. We are going to revert to the previous case. We have that $(y,r) = d^*_{-z}(b,a)$, where $(b,a) \in \partial f(\beta(0,t))$ (see Lemma 8.3) and $z \neq 0$. We are given also that the sequence $(d_{k'},c_{k'})$ approaches $\beta(0,t)$, with $f(d_{k'},c_{k'})$ approaching $f(\beta(0,t))$, and the sequence of regular subgradients $(b_{k'},a_{k'})$ approaches (b,a) and is such that $(b_{k'},a_{k'}) \in \hat\partial f(d_{k'},c_{k'})$ (by Lemma 8.3 again). The rest is analogous to the previous case.

Case 2.c. Suppose finally that there is an infinite subsequence k' such that $c_{k'} = d_{k'}$ for all k'. Then $d_z^*(c_{k'},d_{k'})$ approaches $d_z^*(\beta(0,t)) = (0,t)$, with $f(c_{k'},d_{k'}) = (f\circ\beta)(d_z^*(c_{k'},d_{k'}))$ approaching $f(\beta(0,t)) = (f\circ\beta)(0,t)$ and regular subgradients $(a_{k'},b_{k'}) \in \hat\partial f(\beta(d_z^*(c_{k'},d_{k'})))$. But then by Theorem 8.5 we have that $d_z^*(a_{k'},b_{k'}) \in \hat\partial(f\circ\beta)(0,\sqrt{2}d_{q'})$. Since $d_z^*(a_{k'},b_{k'})$ approaches $d_z^*(a,b)$, we are done.

The proof of the formulae for the horizon subgradient is analogous. ∎

10. Clarke subgradients: The lower semicontinuous case. A function h is caller *lower semicontinuous* if its epigraph $\operatorname{epi} h = \{(x,\alpha) \in \mathbb{R}^n \times \mathbb{R} | h(x) \leq \alpha\}$ is a closed subset of $\mathbb{R}^n \times \mathbb{R}$. Let $C \subset \mathbb{R}^n$ and $\bar{x} \in C$. A vector $v \in \mathbb{R}^n$ is a *regular normal* to C at $\bar{x}$, written $v \in \hat{N}_C(\bar{x})$, if $\operatorname{limsup}_{x \to \bar{x}} \frac{\langle v, x-\bar{x}\rangle}{|x-\bar{x}|} \leq 0$. It is a *normal vector* to C at $\bar{x}$, written $v \in N_C(\bar{x})$, if there is a sequence of points x^k in C approaching $\bar{x}$ and a sequence of regular normals v^k in $\hat{N}_C(x^k)$ approaching v. The set of *Clarke subgradients* of a function h at $\bar{x}$, $\partial^c h(\bar{x})$, is defined by

$$\partial^c h(\bar{x}) = \{v | (v,-1) \in \operatorname{cl}\operatorname{conv} N_{\operatorname{epi} h}(\bar{x}, h(\bar{x}))\}.$$

It can be shown that if h is locally Lipschitz around $\bar{x}$, then this definition coincides with the definition given in section 6, so there is no danger of confusion; see [20, Theorems 9.13(b) and 8.49].

By [20, Theorem 8.9], if h is lower semicontinuous around $\bar{x}$, the following formula holds:

$$N_{\operatorname{epi} h}(\bar{x}, h(\bar{x})) = \{\lambda(v,-1) \,|\, v \in \partial h(\bar{x}),\ \lambda > 0\} \cup \{(v,0) \,|\, v \in \partial^\infty h(\bar{x})\}.$$

The following lemma can be found in [17, Proposition 2.6]. For an independent proof see [15, Lemma 4.1].

LEMMA 10.1. *If h is lower semicontinuous around $\bar{x}$, we have the representation*

$$\partial^c h(\bar{x}) = \operatorname{cl}(\operatorname{conv} \partial h(\bar{x}) + \operatorname{conv} \partial^\infty h(\bar{x})).$$

In particular, when the cone $\partial^\infty h(\bar{x})$ is pointed we have the simpler representation

$$\partial^c h(\bar{x}) = \operatorname{conv} \partial h(\bar{x}) + \operatorname{conv} \partial^\infty h(\bar{x}).$$

It is easy to see that f is lower semicontinuous if and only if $f \circ \beta$ is lower semicontinuous. Our final result is the following theorem.

THEOREM 10.2. *The Clarke subdifferential of any lower semicontinuous, Lorentz invariant function $f \circ \beta$ at the point (x,t) is given by the following formulae:*

(i) *If $x \neq 0$, then*

$$\partial^c(f \circ \beta)(x,t) = \{d_x^*(a,b) \,|\, (a,b) \in \partial^c f(\beta(x,t))\};$$

(ii) *if $x = 0$, then*

$$\partial^c(f \circ \beta)(0,t) = \{d_z^*(a,b) \,|\, (a,b) \in \partial^c f(\beta(0,t)),\ z \neq 0\}.$$

Proof. Suppose first that $x = 0$. Let $A := \partial f(\beta(x,t))$ and $B := \partial^\infty f(\beta(x,t))$. Using Lemmas 5.4 and 10.1 we get

$$\begin{aligned}
\partial^c(f \circ \beta)(x,t) &= \operatorname{cl}(\operatorname{conv} \partial(f \circ \beta)(x,t) + \operatorname{conv} \partial^\infty(f \circ \beta)(x,t)) \\
&= \operatorname{cl}(\operatorname{conv} \mathcal{D}(A) + \operatorname{conv} \mathcal{D}(B)) \\
&= \operatorname{cl}(\mathcal{D}(\operatorname{conv} A) + \mathcal{D}(\operatorname{conv} B)) \\
&= \operatorname{cl} \mathcal{D}(\operatorname{conv} A + \operatorname{conv} B) \\
&= \mathcal{D}(\operatorname{cl}(\operatorname{conv} A + \operatorname{conv} B)) \\
&= \mathcal{D}(\partial^c f(\beta(x,t))).
\end{aligned}$$

The case $x \neq 0$ is analogous. ∎

Acknowledgments. I am very grateful to an anonymous referee for very careful reading of the manuscript and generously providing many useful comments.

REFERENCES

[1] M. Baes, *Spectral functions on Jordan algebras: Differentiability and convexity properties,* CORE Discussion Paper, 2004. Available online at http://www.core.ucl.ac.be/services/psfiles/dp04/dp2004_16.pdf.
[2] H. H. Bauschke, O. Güler, A. S. Lewis, and H. S. Sendov, *Hyperbolic polynomials and convex analysis*, Canad. J. Math., 53 (2001), pp. 470–488.
[3] J. M. Borwein and A. S. Lewis, *Convex Analysis and Nonlinear Optimization*, 2nd ed., Springer, New York, 2006.
[4] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983.
[5] F. H. Clarke, Yu. S. Ledyaev, R. J. Stern, and P. R. Wolenski, *Nonsmooth Analysis and Control Theory*, Springer-Verlag, New York, 1998.
[6] O. Güler, *Hyperbolic polynomials and interior point methods for convex programming*, Math. Oper. Res., 22 (1997), pp. 350–377.
[7] J.-B. Hiriart-Urruty and A. S. Lewis, *The Clarke and Michel-Penot subdifferentials of the eigenvalues of a symmetric matrix*, Comput. Optim. Appl., 13 (1999), pp. 13–23.
[8] A. S. Lewis, *Convex analysis on the Hermitian matrices*, SIAM J. Optim., 6 (1996), pp. 164–177.
[9] A. S. Lewis, *Derivatives of spectral functions*, Math. Oper. Res., 21 (1996), pp. 576–588.
[10] A. S. Lewis, *The convex analysis of unitarily invariant matrix functions*, J. Convex Anal., 2 (1994), pp. 173–183.
[11] A. S. Lewis, *Nonsmooth analysis of eigenvalues*, Math. Program., 84 (1999), pp. 1–24.
[12] A. S. Lewis, P. A. Parrilo, and M. V. Ramana, *The Lax conjecture is true*, Proc. Amer. Math. Soc., 133 (2005), pp. 2495–2499.
[13] A. S. Lewis and H. S. Sendov, *Self-concordant barriers for hyperbolic means*, Math. Program., 91 (2001), pp. 1–10.
[14] A. S. Lewis and H. S. Sendov, *Nonsmooth analysis of singular values.* I. *Theory*, Set-Valued Anal., 13 (2005), pp. 213–241.
[15] A. S. Lewis and H. S. Sendov, *Nonsmooth analysis of singular values.* II. *Applications*, Set-Valued Anal., 13 (2005), pp. 243–264.
[16] M. S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret, *Applications of second-order cone programming*, Linear Algebra Appl., 284 (1998), pp. 193–228.
[17] B. S. Mordukhovich, *Approximation Methods in Problems of Optimization and Control*, Nauka, Moscow, 1988.
[18] J. Renegar, *Hyperbolic programs and their derivative relaxations*, Found. Comput. Math., (2006), pp. 59–79.
[19] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[20] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
[21] H. S. Sendov, *Variational Spectral Analysis*, Ph.D. Thesis, University of Waterloo, Waterloo, Ontario, Canada, 2000.
[22] A. S. Lewis and H. S. Sendov, *Quadratic expansions of spectral functions*, Linear Algebra Appl., 340 (2002), pp. 97–121.
[23] X. D. Chen, D. Sun, and J. Sun, *Complementarity functions and numerical experiments on some smoothing Newton methods for second-order-cone complementarity problems*, Comput. Optim. Appl., 25 (2003), pp. 39–56.
[24] J. S. Pang, D. Sun, and J. Sun, *Semismooth homeomorphisms and strong stability of semidefinite and Lorentz cone complementarity problems*, Math. Oper. Res., 28 (2003), pp. 39–63.
[25] D. Sun and J. Sun, *Löwner's operator and spectral functions in Euclidean Jordan algebras*, Math. Oper. Res., submitted. http://www.math.nus.edu.sg/~matsundf/SS4_Dec24.pdf.

SIAM J. OPTIM.
Vol. 18, No. 4, pp. 1129–1164

STRUCTURAL TOPOLOGY OPTIMIZATION WITH EIGENVALUES*

WOLFGANG ACHTZIGER† AND MICHAL KOČVARA‡

Abstract. The paper considers different problem formulations of topology optimization of discrete or discretized structures with eigenvalues as constraints or as objective functions. We study multiple-load case formulations of minimum volume or weight, minimum compliance problems, and the problem of maximizing the minimal eigenvalue of the structure, including the effect of nonstructural mass. The paper discusses interrelations of the problems and, in particular, shows how solutions of one problem can be derived from solutions of the others. Moreover, we present equivalent reformulations as semidefinite programming problems with the property that, for the minimum volume and minimum compliance problem, each local optimizer of these problems is also a global one. This allows for the calculation of guaranteed global optimizers of the original problems through the use of modern solution techniques of semidefinite programming. For the problem of maximization of the minimum eigenvalue we show how to verify the global optimality and present an algorithm for finding a tight approximation of a globally optimal solution. Numerical examples are provided for truss structures. Both academic and larger-size examples illustrate the theoretical results achieved and demonstrate the practical use of this approach. We conclude with an extension on multiple nonstructural mass conditions.

Key words. eigenvalue optimization, structural optimization, nonlinear semidefinite programming, vibration of structures

AMS subject classifications. 74H45, 74P05, 74P10, 90C22, 90C90

DOI. 10.1137/060651446

1. Introduction.

The subject of this paper is topology optimization of discrete and discretized structures with consideration of free vibrations of the optimal structure. Maximization of the fundamental eigenvalue of a structure is a classic problem in structural engineering. The (generalized) eigenvalue problem typically reads as

$$K(x)w = \lambda(M(x) + M_0)w,$$

where $K(x)$ and $M(x)$ are symmetric and positive semidefinite matrices that continuously (often linearly) depend on the parameter x. The problem has been extensively treated in the engineering literature since the late 1960s; see the papers [24, 21] and the overview [22] summarizing the early development. See also the recent book [26] for an up-to-date bibliography on this subject.

The key difficulty in optimization of structural eigenvalues is the nondifferentiability of the eigenvalues as functions of the design variable x. Many articles in the engineering literature dealing with highly interesting applications use efficient heuristic solution approaches (e.g., [23]). These heuristics are typically based on application-dependent update schemes of the iterates taking some "sensitivity information" into

*Received by the editors February 2, 2006; accepted for publication (in revised form) March 12, 2007; published electronically October 10, 2007. This work was partially done while the first author was visiting the Department of Mathematics, Technical University of Denmark, and the second author was visiting the Institute for Mathematical Sciences, National University of Singapore, in 2006.

http://www.siam.org/journals/siopt/18-4/65144.html

†Institute of Applied Mathematics, University of Dortmund, Vogelpothsweg 87, 44221 Dortmund, Germany (wolfgang.achtziger@uni-dortmund.de).

‡School of Mathematics, University of Birmingham, Edgbaston, Birmingham, B15 3RU, United Kingdom, and Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Pod vodárenskou věží 4, 18208 Praha 8, Czech Republic (kocvara@maths.bham.co.uk).

account mimicking the first order derivative. There are a only few papers using non-heuristic, advanced calculus [25] or the machinery of nondifferentiable optimization [6, 19] to tackle the problems.

In topology optimization there is another serious difficulty coming into play: the fact that components of x can become zero. Notice that in this point topology optimization is different from classic shape optimization. In shape optimization, the design of a structure (a body) is typically described by a sort of parametrization of its boundary, and for each feasible parameter value the corresponding structure can be analyzed by the standard finite element discretization. In contrast to this, (discrete) topology optimization works with a finite element discretization of the whole design space, while the structure itself is given by a material distribution in this space. This means the shape of the structure, the smoothness of its boundary, the number of holes in the structure, the number of finite elements forming the structure, etc., are not predetermined by the user. This results in a set of feasible designs which is enormously rich; it does not require nor does it impose any preliminary knowledge on the resulting design and thus realizes a "free" optimal design process governed solely by the naturally given constraints and physical laws.

The main mathematical difficulty of (discrete) topology optimization is the correct treatment of finite elements which are "empty," i.e., represent holes in the structure. The corresponding matrices $K(x)$ and $M(x)$ then become singular. It seems that this difficulty has not been studied in full mathematical detail so far. The usual way to avoid this singularity is to add an additional constraint, $x \geq \varepsilon > 0$, solve the problem, and interpret the design variables on the lower bounds as zeros. We show in this article that the smallest eigenvalue may be non-Lipschitz and even discontinuous; in certain cases, this fact may cause serious troubles to the above technique.

The general problem of eigenvalue optimization belongs also to classic problems of linear algebra. When the matrix $M(x) + M_0$ is positive definite for all x, then one can resort to the theory developed for the standard eigenvalue problem; see [15] for an excellent overview. Not many papers studying the dependence of the eigenvalues on a parameter are available for the general case when $M(x) + M_0$ is only positive semidefinite; see, e.g., [4, 27, 29].

We present three different formulations of the structural design problem. In the first one we minimize the volume of the structure subject to equilibrium conditions and compliance constraints. Additionally, we require that the fundamental natural frequency of the optimal structure is bigger than or equal to a certain threshold value. The second formulation is analogous; we just switch the volume and the compliance. In the third formulation we maximize the fundamental frequency, i.e., the minimum eigenvalue of certain generalized eigenvalue problem, subject to equilibrium conditions and constraints on the volume and the compliance. Using the semidefinite programming (SDP) framework, we formulate all three problems in a unified way; while the first two problems lead to linear SDP formulations that were already studied earlier [20, 14], the third problem leads to an SDP with a bilinear matrix inequality (BMI) constraint. To our knowledge, the SDP formulation, however straightforward, has never been used for the numerical solution of the third problem. The reason for this was the lack of available SDP-BMI solvers. We solve the problem using the recently developed code PENBMI [10].

We further analyze the mutual relation of our three problems. We show that the problems are in a certain sense equivalent. More precisely, taking a certain specific solution from the solution set of one problem, we get a solution of another problem

with the same data. It is one of the goals of this paper to show that this equivalence does not hold for an arbitrary solution of the problem; this is also illustrated by several numerical examples. Note that there is a common belief that these problems are equivalent. We show that this is not completely true. The equivalence holds only for simplified problems, for instance, when neglecting one of the constraints and assuming that the design variables are strictly positive; see, e.g., [18, p. 212].

An important property of the SDP reformulations of the minimum volume and minimum compliance problem is that each local minimum of any of these problems is also a global minimum. This is not readily seen from the original problem formulations and brings important information to the designer. For the problem of maximization of the minimum eigenvalue we show how to verify the global optimality and present an algorithm for finding an ε-approximation of a globally optimal solution.

Numerical examples conclude the paper. They illustrate the various formulations and theorems developed in the paper and also demonstrate the solvability of the SDP formulations and thus their practical usefulness.

All formulations and theorems in the presentation are developed problems using the discrete structural models, the trusses. This is to keep the notation fixed and simple. The theory also applies to discretized structures, for instance, to the variable thickness sheet or the free material optimization problems [3].

We use standard notation; in particular the notation "$A \succeq 0$" means that the symmetric matrix A is positive semidefinite, and "$A \succ 0$" means that it is positive definite. For two symmetric matrices A, B the notation "$A \succeq B$" ("$A \succ B$") means that $A - B$ is positive semidefinite (positive definite). The $k \times k$ identity matrix is denoted by $I_{k\times k}$; $\ker(A)$ and $\operatorname{range}(A)$ denote the null space and the range space of a matrix A, respectively.

2. Problem definitions, relations.

2.1. Basic notation, generalized eigenvalues. We consider a general mechanical structure, discrete or discretized by the finite element method. The number of members or finite elements is denoted by m. The structure, yet to be optimized, is represented by the vector $x \in \mathbb{R}^m$, $x \geq 0$, of so-called design variables. The meaning of x_i differs from application to application. In truss topology optimization, x_i may represent the volume of a (potential) bar member or its cross-sectional area. In this situation, the m elements form a grid of so-called potential bars which in the classic literature is called a "ground structure" [7, 9]. In discretized problems of continuum topology optimization, x_i may represent the thickness of a sheet. Here the elements refer to a usual finite element discretization, and we must face the difficulty that some of the elements may have zero thickness. In problems of material optimization, x_i may represent a material parameter. Then $x_i = 0$ refers to an extreme material as, e.g., void. Examples of various applications of this modeling can be found in [3]. The total number of "free" degrees of freedom (i.e., not fixed by Dirichlet boundary conditions) is denoted by n. For a given set of n_ℓ (independent) load vectors

$$f_\ell \in \mathbb{R}^n, \quad f_\ell \neq 0, \qquad \ell = 1, \ldots, n_\ell, \tag{1}$$

the structure should satisfy linear equilibrium equations

$$K(x)u_\ell = f_\ell, \qquad \ell = 1, \ldots, n_\ell. \tag{2}$$

Here $K(x)$ is the stiffness matrix of the structure, depending on the design variable x. We will assume linear dependence of K on x,

$$K(x) = \sum_{i=1}^{m} x_i K_i, \tag{3}$$

with $x_i K_i$ being the element stiffness matrices. Note that the stiffness matrix of element (member) e_i is typically defined as

$$x_i K_i = x_i P_i \widehat{K}_i P_i^T, \tag{4}$$

where $P_i P_i^T$ is a projection from $\mathbb{R}^n$ to the space of element (member) degrees of freedom. In other words, $\widehat{K}_i$ is a matrix localized on the particular element, while K_i lives in the full space $\mathbb{R}^n$. Further,

$$x_i \widehat{K}_i = \int_{e_i} x_i B_i^T E_i B_i \, dV,$$

where the rectangular matrix B_i contains derivatives of shape functions of the respective degrees of freedom, and E_i is a symmetric matrix containing information about material properties. To exclude pathological situations, we assume that

$$f_\ell \in \operatorname{range}\left(\sum_{i=1}^{m} K_i\right) \qquad \forall \ell = 1, \dots, n_\ell, \tag{5}$$

which means that there exists a material distribution $x \geq 0$ that can carry all loads f_ℓ (i.e., there exist corresponding $u_1, \dots, u_\ell$ satisfying (2)).

Similarly to the definition of $K(x)$, the mass matrix $M(x)$ of the structure is assumed to be given as

$$M(x) = \sum_{i=1}^{m} x_i M_i, \quad M_i = P_i \widehat{M}_i P_i^T, \tag{6}$$

with element mass matrices

$$x_i \widehat{M}_i = \rho_i \int_{e_i} x_i N_i^T N_i \, dV\,; \tag{7}$$

here N_i contains the shape functions of the degrees of freedom associated with the ith element and ρ_i is the material density.

As already mentioned above, the design variables $x \in \mathbb{R}^m$ represent, for instance, the thickness, cross-sectional area, or material properties of the element. We will assume that

$$x_i \geq 0, \qquad i = 1, \dots, m\,.$$

Notice that the matrices $\widehat{K}_i$, $\widehat{M}_i$ have the properties $\widehat{K}_i \succeq 0$, $\widehat{M}_i \succ 0$, and thus $K(x) \succeq 0$, $M(x) \succeq 0$ for all $x \geq 0$. From a practical point of view, it is worth noticing that the element matrices K_i and M_i are very sparse with only nonzero elements corresponding to degrees of freedom of the ith element. That means, for each i, the matrices K_i and M_i have the same nonzero structure. The matrices $K(x)$, $M(x)$, however, may be dense, in general.

We assume that the discretized structure is connected and the boundary conditions are such that $K(e) \succ 0$ and $M(e) \succ 0$, where e is the vector of all ones. The latter condition simply excludes rigid body movement for any $x > 0$.

In what follows, we will sometimes collect the displacement vectors $u_1, \ldots, u_{n_\ell}$ for all the load cases in one vector,

$$u = (u_1^T, \ldots, u_{n_\ell}^T)^T \in \mathbb{R}^{n \cdot n_\ell},$$

for simplification of the notation.

In this paper we do not rely on any other properties of stiffness and mass matrices than those outlined above. Therefore, the problem formulations and the conclusions apply to a broad class of problems, e.g., to the variable thickness sheet problem or the free material optimization problem [3]. For the sake of transparency, however, we concentrate on a particular class of discrete structures, namely, trusses. A truss is an assemblage of pin-jointed uniform straight bars. The bars are subjected to only axial tension and compression when the truss is loaded at the joints. With a given load and a given set of joints at which the truss is fixed, the goal of the designer is to find a truss that is as light as possible and satisfies the equilibrium conditions. In the simplest, yet meaningful, approach, the number of the joints (nodes) and their position are kept fixed. The design variables are the bar volumes, and the only constraints are the equilibrium equation and an upper bound on the weighted sum of the displacements of loaded nodes, so-called *compliance*. Recently, this model (or its equivalent reformulations) has been extensively analyzed in the mathematical and engineering literature (see, e.g., [1, 3] and the references therein).

In this article, we will additionally consider free vibrations of the optimal structure. The free vibrations are the eigenvalues of the generalized eigenvalue problem

$$K(x)w = \lambda(M(x) + M_0)w\,. \tag{8}$$

The matrix $M_0 \in \mathbb{R}^{n \times n}$ is assumed to be symmetric and positive semidefinite. It denotes the mass matrix of a given nonstructural mass ("dead load"). For the sake of completeness, the choice $M_0 = 0$ is possible and will be treated in more detail below.

In what follows, we use the notation

$$X := \{x \in \mathbb{R}^m \mid x \geq 0,\ x \neq 0\}\,.$$

As a consequence of the construction of $K(x)$ and $M(x)$ we obtain our first result.

LEMMA 2.1. *For each $x \in X$ it holds that*

$$\ker(M(x) + M_0) \subseteqq \ker(K(x))\,.$$

Proof. Let $u \in \mathbb{R}^n$ be in $\ker(M(x) + M_0)$. Then $u^T(M(x) + M_0)u = 0$, that is (see (6)),

$$0 = u^T \left(\sum_{i=1}^m x_i P_i \widehat{M}_i P_i^T + M_0 \right) u = \sum_{i=1}^m x_i (P_i^T u)^T \widehat{M}_i (P_i^T u) + u^T M_0 u\,.$$

Because $\widehat{M}_i \succ 0$ for all i, and because $M_0 \succeq 0$, we conclude that

$$P_i^T u = 0 \qquad \forall i \text{ such that } x_i > 0.$$

Hence, by the definition of $K(x)$ and by (4),

$$K(x)u = \sum_{i=1}^{m} x_i K_i u = \sum_{i=1}^{m} x_i P_i \widehat{K}_i P_i^T u = \sum_{i:\, x_i \neq 0} x_i P_i \widehat{K}_i P_i^T u = 0,$$

and the proof is complete. □

We now want to define a function $\lambda_{\min}$ as the smallest eigenvalue λ of problem (8) for a given structure represented by $x \in X$. Before doing that, we mention the following dilemma in the generalized eigenvalue problem (8). If $x \in X$ is fixed and $(\lambda, w) \in \mathbb{R} \times \mathbb{R}^n$ is a solution of (8) with $w \neq 0$ but $w \in \ker(M(x) + M_0)$, then Lemma 2.1 shows that also $K(x)w = 0$. Hence (μ, w) is also a solution of (8) for arbitrary $\mu \in \mathbb{R}$. In this situation we say that this eigenvalue is *undefined*; otherwise it is *well defined*. Because undefined eigenvalues are meaningless from the engineering point of view, we want to exclude them from our considerations. This leads to the following definition.

DEFINITION 2.2. *For any $x \in X$, let $\lambda_{\min}(x)$ denote the smallest ined eigenvalue of* (8)*, i.e.,*

$$\lambda_{\min}(x) = \min\{\lambda \mid \exists w \in \mathbb{R}^n : \text{ (8) holds for } (\lambda, w) \text{ and } w \notin \ker(M(x) + M_0)\}.$$

This defines a function $\lambda_{\min} : X \longrightarrow \mathbb{R} \cup \{+\infty\}$.

The next proposition collects basic properties of $\lambda_{\min}(\cdot)$.

PROPOSITION 2.3.

(a) *$\lambda_{\min}(\cdot)$ is finite and nonnegative on X.*

(b) *For all $x \in X$,*

$$\lambda_{\min}(x) = \inf_{u:\,(M(x)+M_0)u \neq 0} \frac{u^T K(x) u}{u^T (M(x) + M_0) u}.$$

(c) *For all $x \in X$,*

$$\lambda_{\min}(x) = \sup\{\lambda \mid K(x) - \lambda(M(x) + M_0) \succeq 0\}.$$

(d) *$\lambda_{\min}(\cdot)$ is upper semicontinuous on X.*

(e) *Let $\varepsilon > 0$ be fixed. Then $\lambda_{\min}(\cdot)$ is continuous on $X_\varepsilon := \{x \in \mathbb{R}^m \mid x \geq \varepsilon > 0\}$.*

(f) *$-\lambda_{\min}(\cdot)$ is quasi-convex on X.*

A shorthand version of a proof of this proposition says that in Definition 2.2 we consider a naturally reduced eigenvalue problem which is defined on the space $\ker(A)^\perp$. One can then rely on classic theorems on generalized eigenvalue problems with positive definite matrices. A detailed proof of Proposition 2.3 can be found in Appendix A.1.

For a general $x \in X$ we cannot obtain more than upper semicontinuity of $\lambda_{\min}(\cdot)$ (see Proposition 2.3(d)). The following example shows that $\lambda_{\min}(\cdot)$ may be discontinuous at the boundary of X, when certain components of x are equal to zero.

Example 2.4. Consider the truss depicted in Figure 1. Let the truss be symmetric with respect to (w.r.t.) its horizontal axis, so consider only two design variables, x_1 and x_2. The corresponding stiffness and mass matrix have the following form (where rounded values are displayed for better illustration):

$$K(x) = \begin{pmatrix} x_1 \cdot 2 & 0 & 0 & 0 \\ 0 & x_1 \cdot 2 & 0 & 0 \\ 0 & 0 & x_2 \cdot 1.28 & 0 \\ 0 & 0 & 0 & x_2 \cdot 0.32 \end{pmatrix},$$

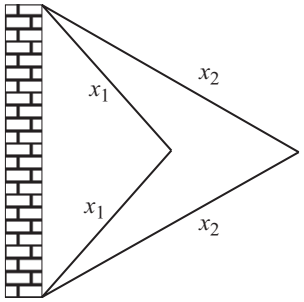


FIG. 1. *Example showing possible discontinuity of* $\lambda_{\min}$.

$$M(x) = \begin{pmatrix} x_1 \cdot 2.83 & 0 & 0 & 0 \\ 0 & x_1 \cdot 2.83 & 0 & 0 \\ 0 & 0 & x_2 \cdot 4.47 & 0 \\ 0 & 0 & 0 & x_2 \cdot 4.47 \end{pmatrix}.$$

The corresponding (unordered) eigenvalues are

$$(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = \Big(\frac{2}{2.83}\frac{x_1}{x_1},\ \frac{2}{2.83}\frac{x_1}{x_1},\ \frac{1.28}{4.47}\frac{x_2}{x_2},\ \frac{0.32}{4.47}\frac{x_2}{x_2} \Big).$$

The function $\lambda_{\min}$ then has the following values:

$$\lambda_{\min}(x) = \frac{0.32}{4.47} \approx 0.07 \quad \text{for } x_2 > 0,$$
$$\lambda_{\min}(x) = \frac{2}{2.83} \approx 0.71 \quad \text{for } x_2 = 0$$

and is thus discontinuous at $x_2 = 0$. The reason for the discontinuity lies in the fact that when $x_2 = 0$ the eigenvalue $\frac{0.32}{4.47}\frac{x_2}{x_2}$ becomes undefined and $\lambda_{\min}$ "jumps" to what was before the second smallest eigenvalue.

Remark 2.5. Example 4.5 will indicate that $\lambda_{\min}(\cdot)$ may not even be Lipschitz continuous near the boundary of X.

2.2. The original formulations. We first give three formulations of the truss design problem that are well known in the engineering literature. These formulations are obtained by just "writing down" the primal requirements and natural constraints.

The minimum volume problem. In the traditional formulation of the truss topology problem, one minimizes the volume or weight of the truss subject to equilibrium conditions and constraints on the smallest eigenfrequency. As mentioned earlier, the meaning of x_i depends on the considered application. Throughout the paper the term $\sum x_i$ is referred to as the "volume" of the structure although its particular interpretation may be different. The term "volume" refers to the truss topology problem where x_i denotes the bar volume of the ith (potential) bar. If $\rho_i > 0$ denotes the density of the material used in this bar, then the substitutions $x_i' := \rho_i x_i$, $K_i' := \frac{1}{\rho_i} K_i$, $M_i' := \frac{1}{\rho_i} M_i$ for all i transform volume into weight. Hence, the following minimum volume problem (and, analogously, all other minimum volume problems in the paper) may be equivalently interpreted as a minimum weight problem, which is of paramount practical interest. In problems of material optimization, x_i may represent a material constant such as Young's modulus of the material in the ith element. In this case $\sum x_i$

represents a simple measure of the total stiffness of the structure. Depending on the precise mechanical meaning of the variables x_i, the matrices K_i and M_i eventually must be scaled by positive constants. Since this does not affect the mathematical properties of these matrices needed in this paper, these modifications are ignored in the following:

$$
\begin{aligned}
(\mathrm{P}_{\mathrm{vol}}) \qquad & \min_{x\in\mathbb{R}^m,\, u\in\mathbb{R}^{n\cdot n_\ell}} \sum_{i=1}^m x_i \\
& \text{subject to } \left(\sum_{i=1}^m x_i K_i\right) u_\ell = f_\ell, \quad \ell = 1,\dots,n_\ell, \\
& \qquad f_\ell^T u_\ell \le \overline{\gamma}, \quad \ell = 1,\dots,n_\ell, \\
& \qquad x_i \ge 0, \quad i = 1,\dots,m, \\
& \qquad \lambda_{\min}(x) \ge \overline{\lambda}\,.
\end{aligned}
$$

Here $\overline{\gamma}$ is a given upper bound on the compliance of the optimal structure and $\overline{\lambda} > 0$ is a given threshold eigenvalue. The simultaneous consideration of constraints on compliance and minimum eigenvalue is useful from a practical point of view. Suitable choices for $\overline{\gamma}$ and $\overline{\lambda}$ are often directly given by technical requirements on the considered application. If a compliance restriction shall not be considered, then $\overline{\gamma}$ formally may be chosen as a very big value. The objective function of this problem is the function $(x,u) \mapsto \sum x_i$. Notice that the eigenvalue constraint is discontinuous (see Example 2.4); this (and not only this) makes the problem rather difficult.

The minimum compliance problem. In this formulation one minimizes the worst-case compliance (maximizes the stiffness) of the truss subject to equilibrium conditions and constraints on the minimum eigenfrequency:

$$
\begin{aligned}
(\mathrm{P}_{\mathrm{compl}}) \qquad & \min_{x\in\mathbb{R}^m, u\in\mathbb{R}^{n\cdot n_\ell}} \max_{1\le\ell\le n_\ell} f_\ell^T u_\ell \\
& \text{subject to } \left(\sum_{i=1}^m x_i K_i\right) u_\ell = f_\ell, \quad \ell = 1,\dots,n_\ell, \\
& \qquad \sum_{i=1}^m x_i \le \overline{V}, \\
& \qquad x_i \ge 0, \quad i = 1,\dots,m, \\
& \qquad \lambda_{\min}(x) \ge \overline{\lambda}\,.
\end{aligned}
$$

Here $\overline{V} > 0$ is an upper bound on the volume of the optimal structure and, again, $\overline{\lambda} > 0$ is a given threshold eigenvalue. For this problem, the objective function is the nonsmooth function $(x,u) \mapsto \max_{1\le\ell\le n_\ell} f_\ell^T u_\ell$. Again, notice that the eigenvalue constraint is not continuous. Moreover, the volume restriction may be formally skipped by choosing $\overline{V}$ very large.

The problem of maximizing the minimal eigenvalue. Here we want to maximize the smallest eigenvalue of (8) subject to equilibrium conditions and constraints on the compliance and volume. Maximization of the smallest eigenfrequency

is of paramount importance in many industrial application, e.g., in civil engineering:

$$
\begin{aligned}
(\mathrm{P}_{\mathrm{eig}}) \qquad & \max_{x\in\mathbb{R}^m, u_\ell\in\mathbb{R}^n} \lambda_{\min}(x) \\
& \text{subject to } \left(\sum_{i=1}^m x_i K_i\right) u_\ell = f_\ell, \quad \ell = 1,\dots,n_\ell, \\
& \qquad f_\ell^T u_\ell \le \overline{\gamma}, \quad \ell = 1,\dots,n_\ell, \\
& \qquad \sum_{i=1}^m x_i \le \overline{V}, \\
& \qquad x_i \ge 0, \quad i = 1,\dots,m\,.
\end{aligned}
$$

Here the objective function is $(x,u) \mapsto \lambda_{\min}(x)$, which is a possibly discontinuous function. This discontinuity is the reason that a standard perturbation approach widely used by practitioners for the solution of $(\mathrm{P}_{\mathrm{eig}})$ may fail. If, with some small $\epsilon > 0$, the nonnegativity constraints are replaced by the constraints $x_i \ge \varepsilon$ for all i, and if x_ε^* denotes a solution of this perturbed problem (together with some u_ε^*), then x_ε^* may not converge to some solution x^* of the unperturbed problem (see Example 2.4 above).

We mention that each of the above three problems has already been considered in the literature with more or less small modifications, and that all problems find valuable interest in practical applications (see [22, 26, 15]). To the best of our knowledge, however, a rigorous treatment of these problems in the situation of positive semidefinite matrices K and M (i.e., permitting $x_i = 0$ for some i, as needed in topology optimization) has not been considered so far.

2.3. Interrelations of original formulations for $M_0 = 0$. In this section we study relations of the three problems $(\mathrm{P}_{\mathrm{vol}})$, $(\mathrm{P}_{\mathrm{compl}})$, and $(\mathrm{P}_{\mathrm{eig}})$ when $M_0 = 0$. These relations are directly given by rescaling arguments but will also appear as special cases of problems with arbitrary M_0 treated in the next section. Note that in the following theorems we do not discuss the *existence* of solutions. Instead, we discuss their interrelations when existence is guaranteed. We start with an auxiliary result.

LEMMA 2.6. *Let $(x,u) \in \mathbb{R}^m \times \mathbb{R}^{n\cdot n_\ell}$, $x \ge 0$, satisfy the equilibrium condition*

$$K(x)u_\ell = f_\ell \tag{9}$$

for some load vector f_ℓ. Then $f_\ell^T u_\ell > 0$ and $\sum_{i=1}^m x_i > 0$.

Proof. Because each of the matrices K_i is symmetric and positive semidefinite, it is clear that $f_\ell^T u_\ell = u_\ell^T K(x) u_\ell \ge 0$. Assume that $f_\ell^T u_\ell = 0$. Then $u_\ell^T K(x) u_\ell = 0$, and simple linear algebra shows that

$$K(x)u_\ell = 0_{\mathbb{R}^n}\,. \tag{10}$$

Equation (10), however, is a contradiction of the assumptions (9) and (1). If $\sum_{i=1}^m x_i = 0$, then $x = 0$, and the contradiction to (9) and (1) is obvious. ∎

Next we observe that the function $\lambda_{\min}(\,.\,)$ is independent of scaling of the structure, provided $M_0 = 0$. Recall that $\lambda_{\min}(x)$ is a well-defined nonnegative number for any $x \in X$ (see Proposition 2.3(a)).

LEMMA 2.7. *Let $M_0 = 0$ and $x \ge 0$ be any vector. Then*

$$\lambda_{\min}(\mu x) = \lambda_{\min}(x) \qquad \forall \mu > 0.$$

Proof. Because $K(\cdot)$ and $M(\cdot)$ are linear functions, the eigenvalue equation $K(\mu x)v = \lambda M(\mu x)v$ is equivalent to $K(x)v = \lambda M(x)v$ for all $\mu > 0$. ∎

We first show that each solution of $(\mathrm{P}_{\mathrm{vol}})$ immediately leads to a solution of $(\mathrm{P}_{\mathrm{compl}})$.

THEOREM 2.8. *Let $M_0 = 0$ and (x^*, u^*) be a solution of $(\mathrm{P}_{\mathrm{vol}})$.*

(a) *Then $\max_{1\le\ell\le n_\ell} f_\ell^T u_\ell^* = \overline{\gamma}$.*

(b) *Put $\overline{V} := \sum_{i=1}^m x_i^*$ in problem $(\mathrm{P}_{\mathrm{compl}})$ and copy the value of $\overline{\lambda}$ from problem $(\mathrm{P}_{\mathrm{vol}})$. Then (x^*, u^*) is optimal for $(\mathrm{P}_{\mathrm{compl}})$ with optimal objective function value $\overline{\gamma}$.*

Proof. For the proof of (a), denote

$$\gamma^* := \max_{1\le\ell\le n_\ell} f_\ell^T u_\ell^*.$$

We must show that $\gamma^* = \overline{\gamma}$. Due to Lemma 2.6 we have

$$\gamma^* > 0 \qquad \text{and} \qquad V^* := \sum x_i^* > 0.$$

Consider the couple

$$(\tilde{x}^*, \tilde{u}^*) := \Big(\frac{\gamma^*}{\overline{\gamma}}\, x^*,\ \frac{\overline{\gamma}}{\gamma^*}\, u^*\Big);$$

by the definition of γ^* we obtain

$$f_\ell^T \tilde{u}_\ell^* = \frac{\overline{\gamma}}{\gamma^*} f_\ell^T u_\ell^* \le \frac{\overline{\gamma}}{\gamma^*}\,\gamma^* = \overline{\gamma} \qquad \forall \ell = 1, \dots, n_\ell, \tag{11}$$

and, obviously,

$$\left(\sum_{i=1}^m \tilde{x}_i^* K_i\right) \tilde{u}_\ell^* = \frac{\gamma^*\overline{\gamma}}{\overline{\gamma}\gamma^*} \left(\sum_{i=1}^m x_i^* K_i\right) u_\ell^* = f_\ell \qquad \forall \ell = 1, \dots, n_\ell.$$

This, together with Lemma 2.7, shows that $(\tilde{x}^*, \tilde{u}^*)$ is feasible for $(\mathrm{P}_{\mathrm{vol}})$. Hence optimality of (x^*, u^*) in $(\mathrm{P}_{\mathrm{vol}})$ yields

$$V^* \le \sum_{i=1}^m \tilde{x}_i^* = \frac{\gamma^*}{\overline{\gamma}} \sum_{i=1}^m x_i^* = \frac{\gamma^*}{\overline{\gamma}} V^*.$$

Because $V^* > 0$, this means

$$\overline{\gamma} \le \gamma^*.$$

Equation (11), however, shows that $\gamma^* \le \overline{\gamma}$. All in all, we arrive at $\gamma^* = \overline{\gamma}$, as stated in (a).

Now we prove (b). Due to the choice of $\overline{V}$ it is clear that (x^*, u^*) is feasible for problem $(\mathrm{P}_{\mathrm{compl}})$. Moreover, (a) shows that the corresponding objective function value is $\overline{\gamma}$. Let (x, u) be an arbitrary feasible point of $(\mathrm{P}_{\mathrm{compl}})$. Lemma 2.6 shows that the value $\gamma := \max_{1\le\ell\le n_\ell} f_\ell^T u_\ell$ is positive, and hence the couple

$$(\tilde{x}, \tilde{u}) := \Big(\frac{\gamma}{\overline{\gamma}}\, x, \frac{\overline{\gamma}}{\gamma}\, u\Big)$$

is well defined. As in (a), we conclude that $(\tilde{x}, \tilde{u})$ is feasible for $(\mathrm{P}_{\mathrm{vol}})$. Optimality of (x^*, u^*) in $(\mathrm{P}_{\mathrm{vol}})$ gives

$$\sum_{i=1}^{m} x_i^* \le \sum_{i=1}^{m} \tilde{x}_i = \frac{\gamma}{\overline{\gamma}} \sum_{i=1}^{m} x_i. \tag{12}$$

Now, $\sum_{i=1}^m x_i^* = \overline{V}$ by the definition of $\overline{V}$, and we have $\sum_{i=1}^m x_i \le \overline{V}$ by the feasibility of (x, u) for $(\mathrm{P}_{\mathrm{compl}})$. Hence (12) becomes $\overline{V} \le \frac{\gamma}{\overline{\gamma}} \overline{V}$, which in turn means that $\overline{\gamma} \le \gamma$. Thus we have shown (use (a)) that

$$\max_{1 \le \ell \le n_\ell} f_\ell^T u_\ell^* \le \max_{1 \le \ell \le n_\ell} f_\ell^T u_\ell,$$

i.e., optimality of (x^*, u^*) for problem $(\mathrm{P}_{\mathrm{compl}})$. ∎

The first assertion of the theorem shows that, when $M_0 = 0$, the compliance constraint in $(\mathrm{P}_{\mathrm{vol}})$ is always active for at least one load case. Later we will demonstrate this theorem by means of a numerical example (see Example 4.1).

A completely analogous theorem to Theorem 2.8 can be stated when problems $(\mathrm{P}_{\mathrm{vol}})$ and $(\mathrm{P}_{\mathrm{compl}})$ are interchanged. The proof uses the same arguments and is thus omitted.

THEOREM 2.9. *Let $M_0 = 0$ and let (x^*, u^*) be a solution of $(\mathrm{P}_{\mathrm{compl}})$.*

(a) *Then $\sum_{i=1}^m x_i^* = \overline{V}$.*

(b) *Put $\overline{\gamma} := \max_{1 \le \ell \le n_\ell} f_\ell^T u_\ell^*$ in problem $(\mathrm{P}_{\mathrm{vol}})$ and copy $\overline{\lambda}$ from $(\mathrm{P}_{\mathrm{compl}})$. Then (x^*, u^*) is optimal for $(\mathrm{P}_{\mathrm{vol}})$ with optimal objective function value $\overline{V}$.*

The interrelations of $(\mathrm{P}_{\mathrm{vol}})$ (resp., of $(\mathrm{P}_{\mathrm{compl}})$) and $(\mathrm{P}_{\mathrm{eig}})$ are a bit more cumbersome because the objective function $(\mathrm{P}_{\mathrm{eig}})$ is invariant with respect to scaling, as shown in Lemma 2.7. As a first and simple result, we obtain the following proposition (where all sums run over $i = 1, \ldots, m$).

PROPOSITION 2.10. *Let $M_0 = 0$, and let (x^*, u^*) be a solution of problem $(\mathrm{P}_{\mathrm{eig}})$.*

(a) *Then for each*

$$\mu \in \left[\frac{\sum x_i^*}{\overline{V}}; \frac{\overline{\gamma}}{\max\limits_{1 \le \ell \le n_\ell} f_\ell^T u_\ell^*} \right] \tag{13}$$

the couple $(\frac{1}{\mu} x^*, \mu u^*)$ is also a solution of $(\mathrm{P}_{\mathrm{eig}})$.

(b) *In particular,*

$$\left(\frac{\overline{V}}{\sum x_i^*}\, x^*,\ \frac{\sum x_i^*}{\overline{V}}\, u^* \right)$$

is also a solution of $(\mathrm{P}_{\mathrm{eig}})$ where the volume constraint is attained as an equality.

(c) *Analogously to* (b),

$$\left(\frac{\max\limits_{1 \le \ell \le n_\ell} f_\ell^T u_\ell^*}{\overline{\gamma}}\, x^*,\ \frac{\overline{\gamma}}{\max\limits_{1 \le \ell \le n_\ell} f_\ell^T u_\ell^*}\, u^* \right)$$

is also a solution of $(\mathrm{P}_{\mathrm{eig}})$ where the compliance constraint is attained as an equality for at least one load case ℓ.

Proof. First, feasibility of (x^*, u^*) in $(\mathrm{P}_{\mathrm{eig}})$ and Lemma 2.6 yield

$$0 < \frac{\sum x_i^*}{\overline{V}} \le 1 \le \frac{\overline{\gamma}}{\max\limits_{1\le \ell \le n_\ell} f_\ell^T u_\ell^*},$$

and hence the interval in (13) is well defined and nonempty. Moreover, it is straightforward to see that

$$\sum \frac{1}{\mu} x_i^* \le \overline{V} \qquad \text{and} \qquad f_\ell^T u_\ell^* \le \overline{\gamma} \quad \forall \ell = 1, \dots, n_\ell$$

hold if and only if μ satisfies (13). Thus for each μ from (13), the point $(\frac{1}{\mu}x^*, \mu u^*)$ is feasible in problem $(\mathrm{P}_{\mathrm{eig}})$. Hence Lemma 2.7 shows that it is even an optimal solution. Assertions (b) and (c) are straightforward consequences of (a). ∎

This proposition relies on the fact that, for $M_0 = 0$, $\lambda_{\min}(\cdot)$ is invariant with respect to scaling of the structure. Hence, if either the volume constraint or the compliance constraints are inactive at the optimum, the optimal structure can be scaled without changing the value of the objective function $\lambda_{\min}(\cdot)$. This shows that (for $M_0 = 0$) problem $(\mathrm{P}_{\mathrm{eig}})$ rather looks for an optimal "shape" of the structure independently of the appropriate scaling. Later, in section 4, we will show a numerical example illustrating Proposition 2.10 (see Example 4.3).

2.4. Interrelations of original formulations for arbitrary M_0. In this section, we do not make any restrictions on M_0 apart from the general requirements already mentioned, i.e., that M_0 is symmetric and positive semidefinite. In the following, when relating two different optimization problems, the matrix M_0 is considered to be the same in both problems.

We start with a general result on the relation of optimization problems where the objective function of one problem acts as a constraint of the other one and vice versa. Through this result we will then be able to state all interrelationships of the formulations $(\mathrm{P}_{\mathrm{vol}})$, $(\mathrm{P}_{\mathrm{compl}})$, and $(\mathrm{P}_{\mathrm{eig}})$.

THEOREM 2.11. *Let $Y \subseteq \mathbb{R}^k$ be nonempty, and let the functions $f_1, f_2 : Y \longrightarrow \mathbb{R}$ be given. For $\overline{f}_1, \overline{f}_2 \in \mathbb{R}$ define the two optimization problems*

$$(\mathrm{P}_1[\overline{f}_2]) \qquad \min_{y\in Y}\{\, f_1(y) \mid f_2(y) \le \overline{f}_2 \,\}$$

and

$$(\mathrm{P}_2[\overline{f}_1]) \qquad \min_{y\in Y}\{\, f_2(y) \mid f_1(y) \le \overline{f}_1 \,\}.$$

Let $\overline{f}_2$ be fixed and the set Y_1^* of solutions to problem $(\mathrm{P}_1[\overline{f}_2])$ be nonempty. The optimal function value is denoted by

$$f_1^* := f_1(y^*) \qquad \forall y^* \in Y_1^*.$$

Put

$$f_2^* := \inf\{\, f_2(y^*) \mid y^* \in Y_1^* \,\}, \tag{14}$$

and let the infimum be attained at some $\hat{y}^* \in Y_1^*$. Consider problem $(\mathrm{P}_2[\overline{f}_1])$ with $\overline{f}_1 := f_1^*$.

Then $\hat{y}^*$ is optimal for problem $(\mathrm{P}_2[\overline{f}_1])$ with optimal objective function value f_2^*.

Proof. Optimality, and hence feasibility, of $\hat{y}^*$ for $(\mathrm{P}_1[\overline{f}_2])$ shows that this point is also feasible for $(\mathrm{P}_2[\overline{f}_1])$ due to the definition of $\overline{f}_1 := f_1^*$. By the choice of $\hat{y}^*$, the value of the objective function of $\hat{y}^*$ in $(\mathrm{P}_2[\overline{f}_1])$ is f_2^*. Now, let y be an arbitrary feasible point of $(\mathrm{P}_2[\overline{f}_1])$ with

$$f_2(y) \leq f_2^*. \tag{15}$$

We must prove that $f_2(y) \geq f_2^*$.

First, the choice of $\hat{y}^*$ shows that

$$f_2^* = f_2(\hat{y}^*) \leq \overline{f}_2.$$

Hence, using (15), we see that

$$f_2(y) \leq \overline{f}_2.$$

Thus, due to feasibility of y in $(\mathrm{P}_2[\overline{f}_1])$, it is clear that (x, u) is also feasible for $(\mathrm{P}_1[\overline{f}_2])$. The definition of $\overline{f}_1$ and the optimality of $\hat{y}^*$ for $(\mathrm{P}_1[\overline{f}_2])$ show that

$$\overline{f}_1 = f_1^* = f_1(\hat{y}^*) \leq f_1(y). \tag{16}$$

The feasibility of (x, u) for $(\mathrm{P}_2[\overline{f}_1])$, however, shows that

$$f_1(y) \leq \overline{f}_1,$$

which together with (16) and with the definition $\overline{f}_1 := f_1^*$ proves

$$f_1(y) = f_1^*.$$

We conclude that y is optimal for $(\mathrm{P}_1[\overline{f}_2])$, i.e., $y \in Y_1^*$. Hence, by the definition of f_2^*,

$$f_2(y) \geq f_2^*,$$

and the proof is complete. ∎

Now we collect certain tools which are needed to show that the infimum in (14) is attained in all situations. For this, we define the function

$$\begin{aligned} c: \ \{\, x \in \mathbb{R}^m \mid x \geq 0 \,\} &\longrightarrow \mathbb{R} \cup \{+\infty\}, \\ x &\mapsto \sup_{1 \leq \ell \leq n_\ell} \sup_{u_\ell \in \mathbb{R}^n} \left\{ 2 f_\ell^T u_\ell - u_\ell^T \left(\sum_{i=1}^m x_i K_i \right) u_\ell \right\}. \end{aligned}$$

Obviously, the function c denotes the maximum (over all load cases) of the negative minimum potential energies of the structure x.

PROPOSITION 2.12 (properties of the function c).

(a) *Let $x \geq 0$. Then $c(x) < +\infty$ if and only if there exist "displacement vectors" $u_1, \ldots, u_{n_\ell} \in \mathbb{R}^n$ such that*

$$K(x) u_\ell = f_\ell \qquad \forall \ell = 1, \ldots, n_\ell. \tag{17}$$

(b) *Let $x \geq 0$. If $c(x) < +\infty$, then*

$$c(x) = \max_{1 \leq \ell \leq n_\ell} f_\ell^T u_\ell$$

for all $u_1, \ldots, u_{n_\ell}$ which satisfy (17).

(c) *The function $c(\cdot)$ is finite and continuous on the set $\{x \in \mathbb{R}^m \mid x > 0\}$ and lower semicontinuous (l.s.c.) on $\{x \mid x \geq 0\}$, i.e.,*

$$\liminf_{\substack{x \to \bar{x} \\ x \geq 0}} c(x) \geq c(\bar{x}), \qquad \bar{x} \geq 0.$$

Proof. All assertions were proved in [2]. Assertions (a) and (b), however, are easily deduced from the necessary and sufficient conditions of the inner sup-problems over u_ℓ and from the fact that a convex quadratic function is unbounded if and only if it does not possess a stationary point. Concerning (c), we mention that the finiteness of c on $\{x \mid x > 0\}$ is based on assumption (5), and that c possesses much stronger continuity properties than just being l.s.c. on $\{x \mid x \geq 0\}$ (see [2]). ∎

For simplification of notation, we define

$$\mathrm{vol}(x) := \sum_{i=1}^{m} x_i$$

for $x \in \mathbb{R}^m$, $x \geq 0$. Moreover, we define

$$\mathcal{S}^*_{\mathrm{vol}}, \mathcal{S}^*_{\mathrm{compl}}, \mathcal{S}^*_{\mathrm{eig}} \subset \{x \in \mathbb{R}^m \mid x \geq 0\} \times \mathbb{R}^{n \cdot n_\ell}$$

as the solution sets of problems ($\mathrm{P_{vol}}$), ($\mathrm{P_{compl}}$), and ($\mathrm{P_{eig}}$), respectively. Notice that these sets may well be empty.

Our first result based on Theorem 2.11 relates problem ($\mathrm{P_{vol}}$) with problems ($\mathrm{P_{compl}}$) and ($\mathrm{P_{eig}}$), respectively.

THEOREM 2.13. *Let $\mathcal{S}^*_{\mathrm{vol}}$ be nonempty. Denote the optimal function value of problem ($\mathrm{P_{vol}}$) by V^*, i.e.,*

$$V^* := \sum_{i=1}^{m} x_i^* \qquad \forall (x^*, u^*) \in \mathcal{S}^*_{\mathrm{vol}}.$$

Put

$$\gamma^* := \inf \left\{ \max_{1 \leq \ell \leq n_\ell} f_\ell^T u_\ell^* \,\middle|\, (x^*, u^*) \in \mathcal{S}^*_{\mathrm{vol}} \right\} \tag{18}$$

and

$$\lambda^* := \sup \left\{ \lambda_{\min}(x^*) \,\middle|\, (x^*, u^*) \in \mathcal{S}^*_{\mathrm{vol}} \right\}. \tag{19}$$

Then the following assertions hold:

(a) *The infimum in (18) is attained at some $(\hat{x}^*, \hat{u}^*) \in \mathcal{S}^*_{\mathrm{vol}}$. Moreover, with $\overline{V} := V^*$, and with $\overline{\lambda}$ copied from problem ($\mathrm{P_{vol}}$), the point $(\hat{x}^*, \hat{u}^*)$ is optimal for problem ($\mathrm{P_{compl}}$) with optimal objective function value γ^*.*

(b) *The supremum in (19) is attained at some $(\tilde{x}^*, \tilde{u}^*) \in \mathcal{S}^*_{\mathrm{vol}}$. Moreover, with $\overline{V} := V^*$, and with $\overline{\gamma}$ copied from problem ($\mathrm{P_{vol}}$), the point $(\tilde{x}^*, \tilde{u}^*)$ is optimal for problem ($\mathrm{P_{eig}}$) with optimal objective function value λ^*.*

Proof. Consider the set

$$\mathcal{X}^*_{\mathrm{vol}} := \{\, x^* \mid (x^*, u^*) \in \mathcal{S}^*_{\mathrm{vol}} \}.$$

Using Proposition 2.12(a) and (b) it is easy to see that

$$\mathcal{X}^*_{\mathrm{vol}} = \left\{ x \geq 0 \,\middle|\, \mathrm{vol}(x) = V^*,\ c(x) \leq \overline{\gamma},\ \lambda_{\min}(x) \geq \overline{\lambda} \right\}. \tag{20}$$

Because $x \geq 0$ and $\mathrm{vol}(x) = V^*$ for all $x \in \mathcal{X}^*_{\mathrm{vol}}$, the set $\mathcal{X}^*_{\mathrm{vol}}$ is bounded. Moreover, because $\mathrm{vol}(\cdot)$ is continuous, $\lambda_{\min}(\cdot)$ is upper semicontinuous (u.s.c.) (see Proposition 2.3(d)) and $c(\cdot)$ is l.s.c. (see Proposition 2.12(c)), the set $\mathcal{X}^*_{\mathrm{vol}}$ is closed. All in all, $\mathcal{X}^*_{\mathrm{vol}}$ is a compact set.

We first prove (a). Proposition 2.12(a) and (b) show that

$$\gamma^* = \inf\{\, c(x) \mid x \in \mathcal{X}^*_{\mathrm{vol}} \,\} \tag{21}$$

and that the infimum in (18) is attained if and only if the infimum in (21) is attained. The latter, however, is straightforward because $c(\cdot)$ is an l.s.c. function, and $\mathcal{X}^*_{\mathrm{vol}}$ is a compact set (each l.s.c. function attains its infimum on a compact set; see, e.g., [17, Thm. 2.13.1]). The rest of the assertion follows directly from Theorem 2.11 with the settings

$$\begin{aligned} Y := \{\, (x,u) \in \mathbb{R}^m \times \mathbb{R}^{n\cdot n_\ell} \mid\ & K(x)u_\ell = f_\ell, && \ell = 1,\dots,n_\ell, \\ & x_i \geq 0, && i = 1,\dots,m, \\ & \lambda_{\min}(x) \geq \overline{\lambda}\}, \end{aligned}$$

$$f_1(x,u) := \mathrm{vol}(x), \qquad f_2(x,u) := c(x), \qquad \overline{f}_2 := \overline{\gamma}, \qquad \overline{f}_1 := V^*.$$

The proof of (b) is analogous. We have to show that the supremum

$$\lambda^* = \sup\{\, \lambda_{\min}(x) \mid x \in \mathcal{X}^*_{\mathrm{vol}} \,\}$$

is attained at some $\tilde{x}^*$. This is the case because $\lambda_{\min}(\cdot)$ is u.s.c. (see Proposition 2.3(d)) and $\mathcal{X}^*_{\mathrm{vol}}$ is compact (see above). Notice that $c(\tilde{x}^*) \leq \overline{\gamma} < +\infty$ (see (20)), and hence corresponding vectors $\tilde{u}^*_1, \dots, \tilde{u}^*_{e_\ell}$ exist by Proposition 2.12(a) and (b) such that $(\tilde{x}^*, \tilde{u}^*)$ is feasible (and optimal) for $(\mathrm{P}_{\mathrm{vol}})$. The rest of assertion (b) follows directly from Theorem 2.11 with the settings

$$\begin{aligned} Y := \{\, (x,u) \in \mathbb{R}^m \times \mathbb{R}^{n\cdot n_\ell} \mid\ & K(x)u_\ell = f_\ell, && \ell = 1,\dots,n_\ell, \\ & x_i \geq 0, && i = 1,\dots,m, \\ & f_\ell^T u_\ell \leq \overline{\gamma}, && \ell = 1,\dots,n_\ell\}, \end{aligned}$$

$$f_1(x,u) := \mathrm{vol}(x), \qquad f_2(x,u) := -\lambda_{\min}(x), \qquad \overline{f}_2 := -\overline{\lambda}, \qquad \overline{f}_1 := V^*. \qquad \square$$

Theorem 2.13(a) reflects the fact that at some solution (x^*, u^*) of $(\mathrm{P}_{\mathrm{vol}})$ none of the compliance constraints may be satisfied with equality, and hence "postoptimization" in (18) is needed to select a proper solution of $(\mathrm{P}_{\mathrm{vol}})$ to obtain a solution of $(\mathrm{P}_{\mathrm{compl}})$. Hence, in general, it is not true that every solution of $(\mathrm{P}_{\mathrm{vol}})$ is also a solution of $(\mathrm{P}_{\mathrm{compl}})$! The problems are thus not equivalent, as is commonly believed. Theorem 2.13(a) also shows that—with the appropriate settings of $\overline{V}$ and $\overline{\lambda}$—there is always a structure x^* which is optimal for *both* problems at the same time (provided there exists a solution at all).

Analogous comments, of course, can be made for Theorem 2.13(b) concerning solutions of $(\mathrm{P}_{\mathrm{eig}})$. A numerical example illustrating Theorem 2.13 is given in section 4 (Example 4.4).

Theorem 2.13 substantially simplifies in the following special situation.

COROLLARY 2.14. *Let the set* $\mathcal{X}^*_{\mathrm{vol}} = \{x^* \mid (x^*, u^*) \in \mathcal{S}^*_{\mathrm{vol}}\}$ *be a singleton. Then the following assertions hold:*

(a) *Put* $\overline{V} := \mathrm{vol}(x^*)$ *in problem* $(\mathrm{P}_{\mathrm{compl}})$ *and copy the value* $\overline{\lambda}$ *from problem* $(\mathrm{P}_{\mathrm{vol}})$. *Then* (x^*, u^*) *is optimal for problem* $(\mathrm{P}_{\mathrm{compl}})$ *with optimal objective function value* $\max_{1\le\ell\le n_\ell} f_\ell^T u_\ell^*$.

(b) *Put* $\overline{V} := \mathrm{vol}(x^*)$ *in problem* $(\mathrm{P}_{\mathrm{eig}})$ *and copy the value* $\overline{\gamma}$ *from problem* $(\mathrm{P}_{\mathrm{vol}})$. *Then* (x^*, u^*) *is optimal for problem* $(\mathrm{P}_{\mathrm{eig}})$ *with optimal objective function value* $\lambda_{\min}(x^*)$.

Proof. If $\mathcal{X}^*_{\mathrm{vol}} = \{x^*\}$, then the infimum in (18) is attained at any $(x^*, u^*) \in \mathcal{S}^*_{\mathrm{vol}}$ because for each $u^*, \tilde{u}^*$ with $K(x^*)u_\ell^* = K(x^*)\tilde{u}_\ell^* = f_\ell$ for all ℓ the compliance values

$$f_\ell^T u_\ell^* = \tilde{u}_\ell^{*T} K(x^*) u_\ell^* = \tilde{u}_\ell^{*T} f_\ell = f_\ell^T \tilde{u}_\ell^*, \qquad \ell = 1, \dots, n_\ell,$$

are constant. Because $\mathcal{X}^*_{\mathrm{vol}}$ is the singleton x^*, and because $\lambda_{\min}(\cdot)$ does not depend on u^*, it is trivial to see that the supremum in (19) is attained at each $(x^*, u^*) \in \mathcal{S}^*_{\mathrm{vol}}$. Now apply Theorem 2.13. ∎

Remark 2.15. Theorem 2.13(a) generalizes Theorem 2.8(b) of the previous chapter. If $M_0 = 0$ in Theorem 2.13(a), then Theorem 2.8(a) shows that

$$\max_{1\le\ell\le n_\ell} f_\ell^T u_\ell^* = \overline{\gamma} \qquad \forall (x^*, u^*) \in S^*_{\mathrm{vol}}.$$

Hence $\gamma^* = \overline{\gamma}$, and the infimum in (18) is attained at each solution $(x^*, u^*) \in \mathcal{S}^*_{\mathrm{vol}}$. A similar comment cannot be made for Theorem 2.13(b). The setting $M_0 = 0$ does not guarantee that for each solution (x^*, u^*) of $(\mathrm{P}_{\mathrm{vol}})$ the eigenvalue constraint is attained as an equality. This will also be demonstrated by Example 4.4 below. The background lies in the invariance of $\lambda_{\min}(\cdot)$ w.r.t. scaling of the structure; see Lemma 2.7.

Analogously to Theorem 2.13, we may derive solutions of problems $(\mathrm{P}_{\mathrm{vol}})$ and $(\mathrm{P}_{\mathrm{eig}})$, respectively, from solutions of problem $(\mathrm{P}_{\mathrm{compl}})$.

THEOREM 2.16. *Let* $\mathcal{S}^*_{\mathrm{compl}}$ *be nonempty. Denote the optimal function value of problem* $(\mathrm{P}_{\mathrm{compl}})$ *by* γ^*, *i.e.,*

$$\gamma^* := \max_{1\le\ell\le n_\ell} f_\ell^T u_\ell^* \qquad \forall (x^*, u^*) \in S^*_{\mathrm{compl}}.$$

Put

$$V^* := \inf \left\{ \sum_{i=1}^m x_i^* \,\middle|\, (x^*, u^*) \in \mathcal{S}^*_{\mathrm{compl}} \right\} \tag{22}$$

and

$$\lambda^* := \sup \left\{ \lambda_{\min}(x^*) \,\middle|\, (x^*, u^*) \in \mathcal{S}^*_{\mathrm{compl}} \right\}. \tag{23}$$

Then the following assertions hold:

(a) *The infimum in* (22) *is attained at some* $(\hat{x}^*, \hat{u}^*) \in \mathcal{S}^*_{\mathrm{compl}}$. *Moreover, with* $\overline{\gamma} := \gamma^*$, *and with* $\overline{\lambda}$ *copied from problem* $(\mathrm{P}_{\mathrm{compl}})$, *the point* $(\hat{x}^*, \hat{u}^*)$ *is optimal for problem* $(\mathrm{P}_{\mathrm{vol}})$ *with optimal objective function value* V^*.

(b) *The supremum in* (23) *is attained at some* $(\tilde{x}^*, \tilde{u}^*) \in \mathcal{S}^*_{\mathrm{compl}}$. *Moreover, with* $\overline{\gamma} := \gamma^*$, *and with* $\overline{V}$ *copied from problem* $(\mathrm{P}_{\mathrm{compl}})$, *the point* $(\tilde{x}^*, \tilde{u}^*)$ *is optimal for problem* $(\mathrm{P}_{\mathrm{eig}})$ *with optimal objective function value* λ^*.

Proof. We modify the proof of Theorem 2.13. Consider the set

$$\mathcal{X}^*_{\mathrm{compl}} := \{\, x^* \mid (x^*, u^*) \in \mathcal{S}^*_{\mathrm{compl}} \}.$$

In view of Proposition 2.12(a) and (b) it is easy to see that

$$\mathcal{X}^*_{\text{compl}} = \Big\{ x \geq 0 \;\Big|\; \text{vol}(x) \leq \overline{V},\ c(x) = \gamma^*,\ \lambda_{\min}(x) \geq \overline{\lambda} \Big\}. \tag{24}$$

Because γ^* is the optimal objective function value, there is no $x \geq 0$ such that $\text{vol}(x) \leq \overline{V}$, $c(x) < \gamma^*$, and $\lambda_{\min}(x) \geq \overline{\lambda}$. Hence the set $\mathcal{X}^*_{\text{compl}}$ remains unchanged if we change the equality sign in "$c(x) = \gamma^*$" to an inequality sign:

$$\mathcal{X}^*_{\text{compl}} = \Big\{ x \geq 0 \;\Big|\; \text{vol}(x) \leq \overline{V},\ c(x) \leq \gamma^*,\ \lambda_{\min}(x) \geq \overline{\lambda} \Big\}. \tag{25}$$

Because $x \geq 0$ and $\text{vol}(x) \leq \overline{V}$ for all $x \in \mathcal{X}^*_{\text{compl}}$, the set $\mathcal{X}^*_{\text{compl}}$ is bounded. Moreover, each of the functions $\text{vol}(\cdot)$, $-\lambda_{\min}(\cdot)$, and $c(\cdot)$ is l.s.c. (see Propositions 2.3(d) and 2.12(c)). Hence the description (25) shows that $\mathcal{X}^*_{\text{compl}}$ is a closed set, and thus $\mathcal{X}^*_{\text{compl}}$ is compact (notice that the *level line* of an l.s.c. function $f(\cdot)$ for some value α, i.e., the set $\{y \mid f(y) = \alpha\}$, need not be closed, but the *level set* $\{y \mid f(y) \leq \alpha\}$ is always closed).

First we prove (a). Obviously, the infimum in (22) is attained because $\mathcal{X}^*_{\text{compl}}$ is a compact set and $\text{vol}(\cdot)$ is continuous. Now apply Theorem 2.11 with the settings

$$\begin{aligned} Y := \{\, (x,u) \in \mathbb{R}^m \times \mathbb{R}^{n \cdot n_\ell} \mid\ & K(x)u_\ell = f_\ell, && \ell = 1, \ldots, n_\ell, \\ & x_i \geq 0, && i = 1, \ldots, m, \\ & \lambda_{\min}(x) \geq \overline{\lambda} \}, \end{aligned}$$

$$f_1(x,u) := \max_{1 \leq \ell \leq n_\ell} f_\ell^T u_\ell, \qquad f_2(x,u) := \text{vol}(x), \qquad \overline{f}_2 := \overline{V}, \qquad \overline{f}_1 := \gamma^*.$$

The proof of (b) is analogous to that of Theorem 2.13(b). □

The following corollary parallels Corollary 2.14. Its proof is even simpler because neither $\text{vol}(\cdot)$ nor $\lambda_{\min}(\cdot)$ in (22) and (23), respectively, depends on u^*.

COROLLARY 2.17. *Let the set* $\mathcal{X}^*_{\text{compl}} = \{x^* \mid (x^*, u^*) \in \mathcal{S}^*_{\text{compl}}\}$ *be a singleton. Then the following assertions hold:*

(a) *Put* $\overline{\gamma} := \max_{1 \leq \ell \leq n_\ell} f_\ell^T u_\ell^*$ *in problem* (P_{vol}) *and copy the value* $\overline{\lambda}$ *from problem* (P_{compl}). *Then* (x^*, u^*) *is optimal for problem* (P_{vol}) *with optimal objective function value* $\text{vol}(x^*)$.

(b) *Put* $\overline{\gamma} := \max_{1 \leq \ell \leq n_\ell} f_\ell^T u_\ell^*$ *in problem* (P_{eig}) *and copy the value* $\overline{V}$ *from problem* (P_{compl}). *Then* (x^*, u^*) *is optimal for problem* (P_{eig}) *with optimal objective function value* $\lambda_{\min}(x^*)$.

Remark 2.18. Similarly as in Remark 2.15, Theorem 2.16(a) generalizes Theorem 2.9(b). If $M_0 = 0$ in Theorem 2.16(a), then Theorem 2.9(a) shows that

$$\text{vol}(x^*) = \overline{V} \qquad \forall (x^*, u^*) \in S^*_{\text{compl}}.$$

Hence $V^* = \overline{V}$, and the infimum in (22) is attained at each solution $(x^*, u^*) \in \mathcal{S}^*_{\text{compl}}$.

Finally, we may derive solutions of problems (P_{vol}) and (P_{compl}) from solutions of (P_{eig}).

THEOREM 2.19. *Let* $\mathcal{S}^*_{\text{eig}}$ *be nonempty. Denote the optimal function value of problem* (P_{eig}) *by* λ^*, *i.e.,*

$$\lambda^* := \lambda_{\min}(x^*) \qquad \forall (x^*, u^*) \in S^*_{\text{eig}}.$$

Put

$$V^* := \inf \left\{ \sum_{i=1}^{m} x_i^* \,\middle|\, (x^*, u^*) \in \mathcal{S}_{\text{eig}}^* \right\} \tag{26}$$

and

$$\gamma^* := \inf \left\{ \max_{1 \le \ell \le n_\ell} f_\ell^T u_\ell^* \,\middle|\, (x^*, u^*) \in \mathcal{S}_{\text{eig}}^* \right\}. \tag{27}$$

Then the following assertions hold:

(a) *The infimum in* (26) *is attained at some* $(\hat{x}^*, \hat{u}^*) \in \mathcal{S}_{\text{eig}}^*$. *Moreover, with* $\overline{\lambda} := \lambda^*$, *and with* $\overline{\gamma}$ *copied from problem* (P_{eig}), *the point* $(\hat{x}^*, \hat{u}^*)$ *is optimal for problem* (P_{vol}) *with optimal objective function value* V^*.

(b) *The infimum in* (27) *is attained at some* $(\tilde{x}^*, \tilde{u}^*) \in \mathcal{S}_{\text{eig}}^*$. *Moreover, with* $\overline{\lambda} := \lambda^*$, *and with* $\overline{V}$ *copied from problem* (P_{eig}), *the point* $(\tilde{x}^*, \tilde{u}^*)$ *is optimal for problem* (P_{compl}) *with optimal objective function value* γ^*.

Proof. The proof of this theorem is analogous to that of Theorem 2.13 with the role of the functions $\text{vol}(\cdot)$ and $\lambda_{\min}(\cdot)$ interchanged. ∎

For an illustration of this theorem, we refer to Example 4.3. The proof of the following corollary is analogous to that of Corollary 2.14.

COROLLARY 2.20. *Let the set* $\mathcal{X}_{\text{eig}}^* = \{x^* \mid (x^*, u^*) \in \mathcal{S}_{\text{eig}}^*\}$ *be a singleton. Then the following assertions hold:*

(a) *Put* $\overline{\lambda} := \lambda_{\min}(x^*)$ *in problem* (P_{vol}) *and copy the value* $\overline{\gamma}$ *from problem* (P_{eig}). *Then* (x^*, u^*) *is optimal for problem* (P_{vol}) *with optimal objective function value* $\text{vol}(x^*)$.

(b) *Put* $\overline{\lambda} := \lambda_{\min}(x^*)$ *in problem* (P_{compl}) *and copy the value* $\overline{V}$ *from problem* (P_{eig}). *Then* (x^*, u^*) *is optimal for problem* (P_{compl}) *with optimal objective function value* $\max_{1 \le \ell \le n_\ell} f_\ell^T u_\ell^*$.

To conclude this theoretical study of relations of the three original problem formulations, we would like to give a few comments on their practical use. Obviously, a direct implementation of one of Theorems 2.13, 2.16, and 2.19 for numerical purposes is difficult because one would need to know the set of *all* solutions to one of the problems, or one should be able to solve the inf- or sup-problems on the optimal set. There are ways to do this, as have been recently shown in [11]. However, as we will see in section 3, there is no need to proceed from a solution of one (nonlinear!) problem to the solution of some other problem, because global solutions of some of the original problems can be calculated through equivalent (quasi-)convex problem formulations.

2.5. Brief discussion on the variation of M_0. In this section we want to briefly prove what is widely known among practitioners: what happens when the nonstructural mass is changed or even removed? For example, if volume minimization is considered, then a bigger nonstructural mass will generally increase the optimal volume. Similarly, if maximization of the minimal eigenvalue is considered, the removal of the nonstructural mass will generally lead to a smaller minimal eigenvalue. Hence, in this section, we briefly consider the variation of M_0 and use the extended notation (see Proposition 2.3(c))

$$\lambda_{\min}(x, M_0) := \sup\{\lambda \mid K(x) - \lambda(M(x) + M_0) \succeq 0\}. \tag{28}$$

LEMMA 2.21. *Let* $x \ge 0$, *and let* $\widetilde{M}_0, M_0 \in \mathbb{R}^{n \times n}$ *be symmetric with* $\widetilde{M}_0 \succeq M_0 \succeq 0$. *Then* $\lambda_{\min}(x, \widetilde{M}_0) \le \lambda_{\min}(x, M_0)$.

Proof. Put $\tilde{\lambda} := \lambda_{\min}(x, \widetilde{M}_0)$. Then

$$\begin{aligned} 0 \preceq K(x) - \tilde{\lambda}(M(x) + \widetilde{M}_0) &= K(x) - \tilde{\lambda}M(x) - \tilde{\lambda}\widetilde{M}_0 \\ \preceq K(x) - \tilde{\lambda}M(x) - \tilde{\lambda}M_0 &= K(x) - \tilde{\lambda}(M(x) + M_0). \end{aligned}$$

Hence,

$$\tilde{\lambda} \leq \sup\{\lambda \mid K(x) - \lambda(M(x) + M_0) \succeq 0\} = \lambda_{\min}(x, M_0). \qquad \square$$

As a simple conclusion concerning the optimal objective function values of our three problems we obtain the following.

PROPOSITION 2.22. *Consider two problems of type* ($\mathrm{P_{vol}}$) *(resp.,* ($\mathrm{P_{compl}}$) *or* ($\mathrm{P_{eig}}$)*), with the same constraint bounds* $\overline{\gamma}$ *and* $\overline{\lambda}$ *(resp.,* $\overline{V}$ *and* $\overline{\lambda}$*, or* $\overline{V}$ *and* $\overline{\gamma}$*) but with different nonstructural mass matrices* $M_0, \widetilde{M}_0$*, where* $\widetilde{M}_0 \succeq M_0$*. Let both problems possess a solution, and denote the optimal objective function values by* $V^*, \widetilde{V}^*$ *(resp.,* $\gamma^*, \widetilde{\gamma}^*$*, or* $\lambda^*, \widetilde{\lambda}^*$*). Then* $V^* \geq \widetilde{V}^*$ *(resp.,* $\gamma^* \geq \widetilde{\gamma}^*$*, or* $\lambda^* \leq \widetilde{\lambda}^*$*).*

Proof. Consider the pair of minimum volume problems. Notice that each feasible point (x, u) of problem ($\mathrm{P_{vol}}$) with nonstructural mass $\widetilde{M}_0$ is also feasible for the problem with nonstructural mass M_0 due to Lemma 2.21(a). Hence, $\tilde{V}^* \leq V^*$.

The proof for the pair of min-max compliance problems is analogous. For the pair of max-min eigenvalue problems it is even simpler, because the set of feasible points is the same for both problems, and Lemma 2.21(a) applies directly on the objective function values. (Notice that for this type of problem, we are *max*imizing, and thus we have "$\leq$" in the assertion.) $\square$

More detailed results than in the above proposition can hardly be obtained, apart from the effect of simple joint scalings of the bounds $\overline{V}, \overline{\gamma}, \overline{\lambda}$, and M_0. Because the total mass matrix in the problem is $(M(x)+M_0)$, a pure change of only M_0 always has nonlinear impact in the problem, and hence is difficult to describe. As a consequence, the optimal topology changes as well with a change of M_0. Such a numerical example is presented section 4 (see Example 4.6).

3. SDP reformulations. All the original formulations are nonconvex, some even discontinuous. Furthermore, all of them implicitly include the computation of the smallest eigenvalue of (8). Below we give reformulations of problems ($\mathrm{P_{vol}}$), ($\mathrm{P_{compl}}$), ($\mathrm{P_{eig}}$) to problems that are much easier to analyze and solve numerically. All these reformulations have been known. The third one, however, has never been used for the numerical treatment, to our knowledge. We will further use a unified approach to these reformulations that offers a clear look at their mutual relations.

We start with an auxiliary result.

PROPOSITION 3.1. *Let* $x \in \mathbb{R}^m$*,* $x \geq 0$*, and* $\gamma \in \mathbb{R}$ *be fixed, and fix an index* $\ell \in \{1, \ldots, n_\ell\}$*. Then there exists* $u_\ell \in \mathbb{R}^n$ *satisfying*

$$K(x)u_\ell = f_\ell \qquad \text{and} \qquad f_\ell^T u_\ell \leq \gamma$$

if and only if

$$\begin{pmatrix} \gamma & -f_\ell^T \\ -f_\ell & K(x) \end{pmatrix} \succeq 0\,.$$

Proof. Note that $K(x)$ may be singular in our case, so that we cannot directly use the Schur complement theorem. We first write the matrix inequality equivalently as

$$\alpha^2\gamma - 2\alpha f_\ell^T v + v^T K(x) v \geq 0 \quad \forall \alpha \in \mathbb{R},\ \forall v \in \mathbb{R}^n\,. \tag{29}$$

"⇒" As $K(x) \succeq 0$, we know that u_ℓ minimizes the quadratic functional $v \mapsto v^T K(x)v - 2f_\ell^T v$ with the minimal value $-f_\ell^T u_\ell$. Thus

$$v^T K(x)v - 2f_\ell^T v \geq -f_\ell^T u_\ell \geq -\gamma \quad \forall v \in \mathbb{R}^n\,.$$

Using the substitution $v = \sigma w$, $\sigma \in \mathbb{R}$, we can write this as

$$(\sigma w)^T K(x)(\sigma w) - 2f_\ell^T(\sigma w) \geq -\gamma \quad \forall \sigma \in \mathbb{R},\ \forall w \in \mathbb{R}^n;$$

hence

$$w^T K(x)w - \frac{1}{\sigma}2f_\ell^T w \geq -\frac{1}{\sigma^2}\gamma \quad \forall \sigma \in \mathbb{R}\setminus\{0\},\ \forall w \in \mathbb{R}^n,$$

which is just (29) with $\alpha = \frac{1}{\sigma}$.

"⇐" Put $\alpha = 1$; then we get from (29)

$$\gamma - 2f_\ell^T v + v^T K(x)v \geq 0 \quad \forall v \in \mathbb{R}^n,$$

meaning that the corresponding convex quadratic function in v is bounded from below. By this, standard linear algebra shows that this function possesses a global minimizer $u_\ell \in \mathbb{R}^n$, and the necessary optimality condition yields

$$K(x)u_\ell = f_\ell\,.$$

Inserting this into (29) with $\alpha = 1$, we have $\gamma - 2f_\ell^T u_\ell + u_\ell^T f_\ell \geq 0$, that is, $\gamma \geq f_\ell^T u_\ell$, and we are done. □

With this proposition, we immediately get the following reformulations of our three original problems.

The minimum volume problem. In this problem, $\overline{\gamma}$ and $\overline{\lambda}$ are given, and we minimize the upper bound V on the volume:

$$(\mathrm{P}_{\mathrm{vol}}^{\mathrm{SDP}}) \qquad \begin{aligned} \min_{x\in\mathbb{R}^m, V\in\mathbb{R}} \ & V \\ \text{subject to } & \begin{pmatrix} \overline{\gamma} & -f_\ell^T \\ -f_\ell & K(x) \end{pmatrix} \succeq 0, \quad \ell = 1,\dots,n_\ell, \\ & \sum_{i=1}^m x_i \leq V, \\ & x_i \geq 0, \quad i = 1,\dots,m, \\ & K(x) - \overline{\lambda}(M(x) + M_0) \succeq 0\,. \end{aligned}$$

A closely related version of this problem was first formulated and studied by Ohsaki et al. [20]. They considered the problem without the compliance constraint and with positive lower bounds on the design variables x_i. Using SDP duality, global optima of this problem can be easily characterized by necessary and sufficient optimality conditions [8]. Analogous statements can be made for problem $(\mathrm{P}_{\mathrm{vol}}^{\mathrm{SDP}})$ as well.

The minimum compliance problem. Here $\overline{V}$ and $\overline{\lambda}$ are given, and we minimize the upper bound γ on the compliance:

$$
\begin{aligned}
(\mathrm{P}^{\mathrm{SDP}}_{\mathrm{compl}}) \qquad & \min_{x\in\mathbb{R}^m,\gamma\in\mathbb{R}} \gamma \\
& \text{subject to } \begin{pmatrix} \gamma & -f_\ell^T \\ -f_\ell & K(x) \end{pmatrix} \succeq 0, \quad \ell = 1,\dots,n_\ell, \\
& \sum_{i=1}^m x_i \le \overline{V}, \\
& x_i \ge 0, \quad i = 1,\dots,m, \\
& K(x) - \overline{\lambda}(M(x) + M_0) \succeq 0\,.
\end{aligned}
$$

The problem of maximizing the minimal eigenvalue. Now $\overline{\gamma}$ and $\overline{V}$ are given, and λ is the variable. For the sake of a common problem structure in all three formulations, we *minimize* and put a minus in front of the objective function:

$$
\begin{aligned}
(\mathrm{P}^{\mathrm{SDP}}_{\mathrm{eig}}) \qquad & \min_{x\in\mathbb{R}^m,\lambda\in\mathbb{R}} -\lambda \\
& \text{subject to } \begin{pmatrix} \overline{\gamma} & -f_\ell^T \\ -f_\ell & K(x) \end{pmatrix} \succeq 0, \quad \ell = 1,\dots,n_\ell, \\
& \sum_{i=1}^m x_i \le \overline{V}, \\
& x_i \ge 0, \quad i = 1,\dots,m, \\
& K(x) - \lambda(M(x) + M_0) \succeq 0\,.
\end{aligned}
$$

The proof of the following proposition is immediate, and thus is skipped.

PROPOSITION 3.2.

(a) *If* (x^*,u^*) *is a global minimizer of* $(\mathrm{P}_{\mathrm{vol}})$*, then* (x^*,V^*) *is a global minimizer of* $(\mathrm{P}^{\mathrm{SDP}}_{\mathrm{vol}})$*, where* $V^* := \sum x_i^*$*, and the optimal values of both problems coincide.*

(b) *If* (x^*,V^*) *is a global minimizer of* $(\mathrm{P}^{\mathrm{SDP}}_{\mathrm{vol}})$*, then there exists* u^* *such that* (x^*,u^*) *is a global minimizer of* $(\mathrm{P}_{\mathrm{vol}})$*, and the optimal values of both problems coincide.*

Analogous statements hold for the pairs of problems $(\mathrm{P}_{\mathrm{compl}})$*,* $(\mathrm{P}^{\mathrm{SDP}}_{\mathrm{compl}})$ *and* $(\mathrm{P}_{\mathrm{eig}})$*,* $(\mathrm{P}^{\mathrm{SDP}}_{\mathrm{eig}})$*, respectively, where in the latter case the optimal function values coincide up to a sign.*

Note that problems $(\mathrm{P}^{\mathrm{SDP}}_{\mathrm{vol}})$ and $(\mathrm{P}^{\mathrm{SDP}}_{\mathrm{compl}})$ are linear SDPs, while $(\mathrm{P}^{\mathrm{SDP}}_{\mathrm{eig}})$ is an SDP problem with a BMI constraint, i.e., it is generally nonconvex. We should emphasize that, due to the SDP reformulation, the originally discontinuous problems became continuous, a fact of big practical value.

THEOREM 3.3. *Each local minimizer of problem* $(\mathrm{P}^{\mathrm{SDP}}_{\mathrm{vol}})$ *is also a global minimizer. An analogous statement holds for problem* $(\mathrm{P}^{\mathrm{SDP}}_{\mathrm{compl}})$*.*

Proof. Problems $(\mathrm{P}^{\mathrm{SDP}}_{\mathrm{vol}})$ and $(\mathrm{P}^{\mathrm{SDP}}_{\mathrm{compl}})$ are linear SDPs, i.e., convex problems, and the assertions follow. ∎

Needless to say that this theorem is of paramount interest from the practical point of view.

Clearly, a statement similar to Theorem 3.3 does not hold for problem $(\mathrm{P}^{\mathrm{SDP}}_{\mathrm{eig}})$; see Example 2.4, where the function $\lambda_{\min}(\cdot)$ is constant for $x_2 > 0$ and has thus infinitely many local minima which are, however, greater than the global minimum attained at $x_2 = 0$.

We remark, however, that problem $(\mathrm{P}_{\mathrm{eig}}^{\mathrm{SDP}})$ hides a *quasi*-convex structure. To see this, use Definition 2.2 to write problem $(\mathrm{P}_{\mathrm{eig}}^{\mathrm{SDP}})$ in the form

$$\min\{\, -\lambda_{\min}(x) \mid x \in \mathcal{F}\,\} \tag{30}$$

with the feasible set

$$\mathcal{F} := \left\{ x \in \mathbb{R}^m \,\middle|\, x \geq 0;\ \begin{pmatrix} \overline{\gamma} & -f_\ell^T \\ -f_\ell & K(x) \end{pmatrix} \succeq 0,\ \ell = 1, \ldots, n_\ell;\ \sum_{i=1}^m x_i \leq \overline{V} \right\}.$$

Then Proposition 2.3(f) and the fact that the cone of positive semidefinite matrices is convex show that we minimize here a quasi-convex function over a convex feasible set $\mathcal{F}$. This fact might be useful, e.g., for the application of cutting plane algorithms from global optimization. Unfortunately, the function $-\lambda_{\min}(\cdot)$ fails to be strictly quasi-convex, as already explained in Example 2.4.

Formulation (30) of problem $(\mathrm{P}_{\mathrm{eig}}^{\mathrm{SDP}})$ immediately clarifies the existence of solutions.

THEOREM 3.4. *Problem* $(\mathrm{P}_{\mathrm{eig}}^{\mathrm{SDP}})$ *(or, equivalently, problem* $(\mathrm{P}_{\mathrm{eig}})$*) possesses a solution if and only if it possesses feasible points.*

Proof. Consider problem $(\mathrm{P}_{\mathrm{eig}}^{\mathrm{SDP}})$ in the form (30). Since the cone of positive semidefinite matrices is closed, the set $\mathcal{F}$ is compact. Moreover, $0 \notin \mathcal{F}$ due to assumption (1), and hence $(-\lambda_{\min})$ is l.s.c. on $\mathcal{F}$ by Proposition 2.3(d). Each l.s.c. function attains its infimum on a nonempty compact set (see, e.g., [17, Thm. 2.13.1]). ∎

Instead of using methods of global optimization for the calculation of a global minimizer of problem $(\mathrm{P}_{\mathrm{eig}}^{\mathrm{SDP}})$, we may use the quasi-convex formulation (30) and the close relation to the convex problem $(\mathrm{P}_{\mathrm{vol}}^{\mathrm{SDP}})$. In the following we define a bisection technique that is based on a general algorithm for problems with quasi-convex objective function and convex constraints; see, e.g., [5, sect. 4.2.5]. This technique of finding a global solution of $(\mathrm{P}_{\mathrm{eig}}^{\mathrm{SDP}})$ is based on the solutions of a sequence of problems which are of the type $(\mathrm{P}_{\mathrm{vol}}^{\mathrm{SDP}})$. Alternatively, a sequence of problems of the type $(\mathrm{P}_{\mathrm{compl}}^{\mathrm{SDP}})$ can be used (not described here).

The methodology works as follows. We may consider problem $(\mathrm{P}_{\mathrm{vol}}^{\mathrm{SDP}})$ as a kind of feasibility problem for $(\mathrm{P}_{\mathrm{eig}}^{\mathrm{SDP}})$ where $\lambda := \overline{\lambda}$ is fixed. More precisely, by solving the convex problem $(\mathrm{P}_{\mathrm{vol}}^{\mathrm{SDP}})$ we can see whether $(\mathrm{P}_{\mathrm{eig}}^{\mathrm{SDP}})$ possesses feasible points with objective function value $\lambda \geq \overline{\lambda}$ or not. Hence, we may state a simple bisection technique playing with the value of $\overline{\lambda}$. Since the volume constraint of $(\mathrm{P}_{\mathrm{eig}}^{\mathrm{SDP}})$ is not a constraint of $(\mathrm{P}_{\mathrm{vol}}^{\mathrm{SDP}})$, we slightly modify $(\mathrm{P}_{\mathrm{vol}}^{\mathrm{SDP}})$. For fixed $\lambda \geq 0$ and fixed $\delta \geq 0$ consider the following linear SDP:

$$(\mathrm{P}_{\mathrm{vol}}^{\mathrm{SDP}}(\lambda, \delta)) \qquad \min_{x \in \mathbb{R}^m, V \in \mathbb{R}} V$$

$$\text{subject to } \begin{pmatrix} \overline{\gamma} & -f_\ell^T \\ -f_\ell & K(x) \end{pmatrix} \succeq 0, \quad \ell = 1, \ldots, n_\ell,$$

$$\sum_{i=1}^m x_i \leq V,$$

$$V \leq \overline{V},$$

$$x_i \geq 0, \quad i = 1, \ldots, m,$$

$$K(x) - (\lambda + \delta)(M(x) + M_0) \succeq 0\,.$$

In the following, the feasible set of this problem is denoted by

$$\mathcal{F}(\lambda, \delta),$$

for simplicity. Since $(\mathrm{P}^{\mathrm{SDP}}_{\overline{\mathrm{vol}}}(\lambda, \delta))$ is a convex SDP, modern solution procedures are able to recognize whether $\mathcal{F}(\lambda, \delta) = \varnothing$, and we may calculate a feasible point or even a global minimizer if $\mathcal{F}(\lambda, \delta) \neq \varnothing$.

The following proposition gives a tool for the estimation of the (globally) optimal objective function value of problem $(\mathrm{P}^{\mathrm{SDP}}_{\mathrm{eig}})$. Its proof is a simple exercise.

PROPOSITION 3.5. *Let $(\tilde{x}, \lambda)$ be feasible for $(\mathrm{P}^{\mathrm{SDP}}_{\mathrm{eig}})$, and let $(-\lambda^{**})$ denote the (globally) optimal function value of problem $(\mathrm{P}^{\mathrm{SDP}}_{\mathrm{eig}})$. Moreover, let $\delta > 0$ be arbitrary, and consider problem $(\mathrm{P}^{\mathrm{SDP}}_{\overline{\mathrm{vol}}}(\lambda, \delta))$ with the parameters $\overline{\gamma}$ and $\overline{V}$ copied from $(\mathrm{P}^{\mathrm{SDP}}_{\mathrm{eig}})$. Then the following assertions hold:*

(a) *If $\mathcal{F}(\lambda, \delta) \neq \varnothing$, then for each $(x, V) \in \mathcal{F}(\lambda, \delta)$ the point $(x, \lambda + \delta)$ is feasible for $(\mathrm{P}^{\mathrm{SDP}}_{\mathrm{eig}})$, i.e., $-\lambda^{**} \leq -(\lambda + \delta) < -\lambda$.*

(b) *If $\mathcal{F}(\lambda, \delta) = \varnothing$, then $-(\lambda + \delta) < -\lambda^{**} \leq -\lambda$.*

The practical value of this proposition lies in the possibility of improving upper and lower bounds for λ^{**} which can be numerically calculated through solutions (or only feasible points) of convex linear SDPs.

As a preprocessing step, we first calculate initial lower and upper bounds λ_0^L, λ_0^U on λ^{**}. For this, first calculate a feasible point (x, λ) of $(\mathrm{P}^{\mathrm{SDP}}_{\mathrm{eig}})$ and choose arbitrary $\bar{\delta} > 0$. Then find the smallest $k \in \mathbb{N}$ such that $\mathcal{F}(\lambda, 2^k\bar{\delta}) = \varnothing$ by solving $(\mathrm{P}^{\mathrm{SDP}}_{\overline{\mathrm{vol}}}(\lambda, 2^k\bar{\delta}))$ repeatedly. Set $\lambda_0^L := \lambda + 2^{k-1}\bar{\delta}$ and $\lambda_0^U := \lambda + 2^k\bar{\delta}$. Then Proposition 3.5 shows that

$$0 \leq \lambda_0^L \leq \lambda^{**} < \lambda_0^U \ .$$

With these bounds it is easy to construct a bisection-type algorithm which in each step reduces the gap $(\lambda_k^U - \lambda_k^L)$ by a factor of (at least) $\frac{1}{2}$.

ALGORITHM 3.6. Choose an accuracy $\eta > 0$, a feasible point $(\hat{x}, \hat{\lambda})$ for $(\mathrm{P}^{\mathrm{SDP}}_{\mathrm{eig}})$. Put $(x_0, \lambda_0) := (\hat{x}, \hat{\lambda})$, $\delta_0 := \frac{1}{2}(\lambda_0^U - \lambda_0^L)$, and $k := 0$. Go to step 2.

1. Calculate a feasible point [or even a local minimizer] (x_k, λ_k) of $(\mathrm{P}^{\mathrm{SDP}}_{\mathrm{eig}})$ with the additional constraint "$\lambda \geq \lambda_k^L$".
2. If $\lambda_k > \lambda_k^L$, then update λ_k^L by $\lambda_k^L := \lambda_k$.
3. If $\lambda_k^U - \lambda_k^L \leq \eta$, then EXIT with the result $(x^*, \lambda^*) := (x_k, \lambda_k)$.
4. Put $\delta_k := \frac{1}{2}(\lambda_k^U - \lambda_k^L)$, and consider problem $(\mathrm{P}^{\mathrm{SDP}}_{\overline{\mathrm{vol}}}(\lambda_k, \delta_k))$.
   If $\mathcal{F}(\lambda_k, \delta_k) \neq \varnothing$, then:
   4A. Put $\lambda_{k+1}^L := \lambda_k^L + \delta_k$, $k := k + 1$, and go to step 1.
   Otherwise, if $\mathcal{F}(\lambda_k, 2^k\bar{\delta}) = \varnothing$, then:
   4B. Put $\lambda_{k+1}^U := \lambda_k^U - \delta_l$, $k := k + 1$, and go to step 1.

Let $(\mathrm{P}^{\mathrm{SDP}}_{\mathrm{eig}})$ possess a solution (x^{**}, λ^{**}) (see Theorem 3.4). Then it is straightforward to show that Algorithm 3.6 is well defined, and that after each iteration the inequalities $\lambda_k^L \leq \lambda^{**} < \lambda_k^U$ and $\lambda_k^U - \lambda_k^L \leq 2^{-k}(\lambda_0^U - \lambda_0^L)$ hold. Consequently, Algorithm 3.6 terminates after at most $\lceil(\ln(\lambda_0^U - \lambda_0^L) - \ln(\eta))/(\ln(2))\rceil$ iterations. Moreover, at termination, the result (x^*, λ^*) is feasible for $(\mathrm{P}^{\mathrm{SDP}}_{\mathrm{eig}})$ with $\lambda^{**} - \lambda^* \leq \eta$.

Notice that the additional constraint "$\lambda \geq \lambda_k^L$" in step 1 does not cause any trouble but guarantees that $(\lambda_k)_k$ is monotonically increasing. Moreover, the calculation of global minimizers (in step 4A), respectively, local minimizers (in step 1), instead of just feasible points should significantly speed up the algorithm. In this case the update of λ_k^U in step 4B, respectively, of λ_k^L in step 2, may lead to a much bigger

reduction of the gap $\lambda_k^U - \lambda_k^U$. Obviously, step 1 must be carried out in each iteration. Notice also that λ_k^L is increased in step 4A, while it remains untouched in step 4B. Denote by K the number of iterations in which step 4A has been performed. Moreover, if step 4A has been performed in iteration $k-1$, let (x_k, λ_k) in step 1 be a local optimizer. Then, consequently,

$$K \leq \left| \left\{ \lambda \,|\, (x, \lambda) \text{ is a local optimizer of } (\mathrm{P}_{\mathrm{eig}}^{\mathrm{SDP}}) \right\} \right|,$$

i.e., K is limited by the number of levels of the objective function which are attained at a local optimizer. We believe that this cardinality is very small in applications. As an illustration consider Example 2.4 where $K = 2$.

For the numerical treatment of the SDP problems $(\mathrm{P}_{\mathrm{vol}}^{\mathrm{SDP}})$, $(\mathrm{P}_{\mathrm{compl}}^{\mathrm{SDP}})$, $(\mathrm{P}_{\mathrm{eig}}^{\mathrm{SDP}})$ one must resort to methods of semidefinite programming. Such methods, and corresponding codes, are nowadays available for linear SDPs. The limiting factor of these codes is, however, the problem size which, compared to general nonlinear programs, is restricted to problems of medium size. The problem $(\mathrm{P}_{\mathrm{eig}}^{\mathrm{SDP}})$ even requires a method which can deal with BMIs. We will use such a method to solve examples in the next section. It should be noted, however, that algorithms and codes for SDPs with BMIs are on the edge of current research and are not yet standard. Notice that the BMI constraint in any generalized eigenvalue problem (GEVP) has a very simple structure. As above, being quasi-convex, the problem can be solved by a bisection algorithm, i.e., by solving a sequence of linear SDP problems of the same size and structure (again, see [5, sect. 4.2.5]). However, the crucial factor, in our opinion, is the *size* of the problem. To solve one GEVP problem formulated as BMI (for instance, by the code PENBMI), one needs about the same CPU time as to solve one linear SDP problem in the bisection algorithm! Hence, when we want to apply the approach to large-scale problems with thousands of variables (in particular, to problems discretized by the finite element method), the BMI formulation will use only a fraction of CPU time when compared to the bisection approach.

4. Numerical examples. In this chapter we present numerical examples which, on the one hand, will illustrate some of the theoretical results above and, on the other hand, demonstrate the practical use of the SDP problem formulations.

Intentionally, we do not specify realistic physical properties, like material elastic modulus E or density ρ. In all examples, they are assumed to be equal to one. This can be done without any loss of generality or applicability due to the linear dependence of the volume, compliance, and stiffness and mass matrices on x, E, and ρ. In this case, switching from our default values to realistic physical values (say, $E = 2.1 \cdot 10^{11}$ Pa and $\rho = 7.8 \cdot 10^3$ kg/m^3) is a matter of a simple linear scaling of our results. Further, when we speak of an $i \times j$ truss, we have in mind a regular equidistant grid of $i \cdot j$ nodes, i in the horizontal direction and j in the vertical direction. Thus the dimensions of the nodal grid are $(i-1) \times (j-1)$.

The code we have used for the treatment of the SDP formulations is PENBMI, version 2.0 [13]. This code implements the generalized augmented Lagrangian method, as described in [12, 28]. In particular, PENBMI can treat BMIs, as is necessary for problem $(\mathrm{P}_{\mathrm{eig}}^{\mathrm{SDP}})$ [10].

The examples were solved on a Pentium III-M 1GHz PC running Windows 2000. All problems were formulated and solved in MATLAB using the YALMIP parser [16] to PENBMI.

Example 4.1. This example illustrates Theorems 2.8, 2.9, and 2.19 with $M_0 = 0$. Consider a 3×3 truss with all nodes connected by potential bars. The nodes on the

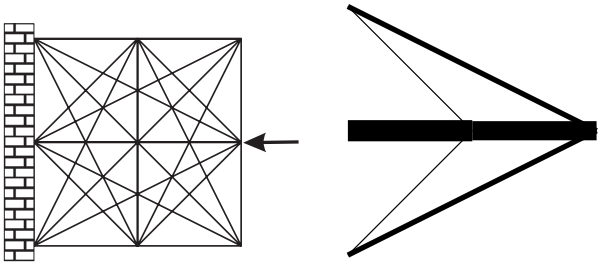


FIG. 2. *A* 3×3 *truss (Example* 4.1*): initial layout and optimal topology.*

left-hand side are fixed in both directions and a horizontal force $(-1, 0)$ is applied at the right-middle node; see Figure 2 (left). No nonstructural mass is considered, i.e., $M_0 = 0$. We consider the minimum volume problem $(\mathrm{P}_{\mathrm{vol}}^{\mathrm{SDP}})$ with $\overline{\gamma} = 1$ and $\overline{\lambda} = 5.0 \cdot 10^{-2}$. PENBMI calculated the (global) optimal solution (x^*, V^*) of this convex problem: the optimal design x^* is shown in Figure 2 (right), while $V^* = 1.20229$. Proposition 3.2(b) shows that there exists u^* such that (x^*, u^*) is optimal for problem $(\mathrm{P}_{\mathrm{vol}})$.

Now consider the minimum compliance problem $(\mathrm{P}_{\mathrm{compl}}^{\mathrm{SDP}})$ with $\overline{V} = 1.20229$ and $\overline{\lambda} = 5.0 \cdot 10^{-2}$. As expected by Proposition 3.2(b) and Theorem 2.9, we obtain the solution (x^*, γ^*) with the same structure x^* as before (Figure 2 (right)), and with $\gamma^* = 1$.

Finally, when solving the problem of maximizing the minimum eigenvalue $(\mathrm{P}_{\mathrm{eig}}^{\mathrm{SDP}})$ with $\overline{V} = 1.20229$ and $\overline{\gamma} = 1$, we again obtain x^* from before, and $\lambda^* = 5.0 \cdot 10^{-2}$. This shows that the value V^* in (26) and the value γ^* in (27) are attained for x^*, because otherwise this would yield a contradiction to Theorem 2.19. The authors believe that in this simple example the solution structure x^* is the unique solution, and thus Corollary 2.20 may be applied.

Example 4.2. In this example, as in Example 4.1 above, we again obtain the same optimal structure for all three problem formulations. Here, however, $M_0 \neq 0$, and thus these coincidences are somewhat unexpected.

We consider the same ground structure, boundary conditions, and external load as in the previous example. In addition, we assign a nonstructural mass of size 10 at the loaded node, i.e., $M_0 \neq 0$; see Figure 3 (left). Consider the minimum volume problem $(\mathrm{P}_{\mathrm{vol}}^{\mathrm{SDP}})$ with $\overline{\gamma} = 1$ and $\overline{\lambda} = 5.0 \cdot 10^{-2}$. Figure 3 (right) shows the optimal design x^*. The corresponding optimal volume is $V^* = 7.10157$.

Now consider the minimum compliance problem $(\mathrm{P}_{\mathrm{compl}}^{\mathrm{SDP}})$ with $\overline{V} = 7.1015$ and $\overline{\lambda} = 5.0 \cdot 10^{-2}$. We obtain the solution (x^*, γ^*) with the same structure x^* as before (Figure 3 (right)), and with $\gamma^* = 1$.

Finally, when solving the problem of maximizing the minimum eigenvalue $(\mathrm{P}_{\mathrm{eig}}^{\mathrm{SDP}})$ with $\overline{V} = 7.1015$ and $\overline{\gamma} = 1$, we again obtain x^* from above, and $\lambda^* = 5.0 \cdot 10^{-2}$. Again, we believe that the solution x^* is unique in each of the three problems. If this is the case, then the equivalence of the results holds by Corollaries 2.14, 2.17, and 2.20.

Example 4.3. This academic example illustrates the possible nonuniqueness of solutions to the problem $(\mathrm{P}_{\mathrm{eig}}^{\mathrm{SDP}})$. Consider a 2×3 ground structure with boundary conditions and load as depicted in Figure 4 (left). Put $M_0 = 0$, $\overline{\gamma} = 10$, and $\overline{V} = 10$. The computed optimal structure x^* is presented in Figure 4 (right); the optimal

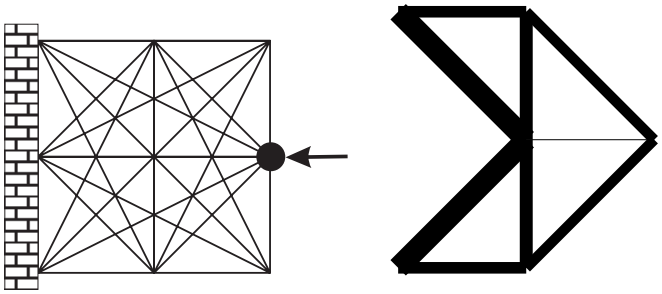


FIG. 3. *A* 3×3 *truss with nonstructural mass (Example* 4.2*): initial layout and optimal topology.*

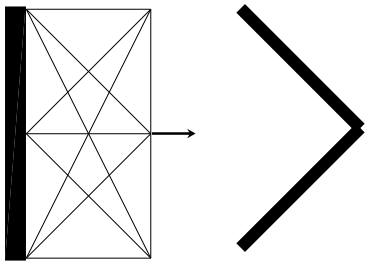


FIG. 4. *Example demonstrating possible nonuniqueness of solution of the* ($\mathrm{P}_{\mathrm{eig}}^{\mathrm{SDP}}$) *problem.*

objective function value of ($\mathrm{P}_{\mathrm{eig}}^{\mathrm{SDP}}$) is $-\lambda^* = -0.70711$, i.e., $\lambda_{\min}(x^*) = 0.70711$. While the volume constraint is active at x^*, the compliance constraint is inactive (more precisely, after calculating some u^* corresponding to x^*, we have $\gamma^* := f^T u^* = 0.1 < \overline{\gamma} = 10$). Proposition 2.10 suggests that if we scale the solution x^* by a certain factor μ, we will still get a solution to our problem. For instance, if we solve the same problem but with $\overline{V} = 1.0$, then we will obtain a solution with the same λ^* and with $\gamma^* = 1.0$, i.e., still within the $\overline{\gamma}$ limits. Table 1 summarizes these numbers. It also presents the results for the case when $M_0 = 10$ (and then Proposition 2.10 does not apply). In this case, the optimal solution is no longer scalable.

Example 4.4. Here we demonstrate the possible nonuniqueness of solutions to the minimum volume problem ($\mathrm{P}_{\mathrm{vol}}$) (or ($\mathrm{P}_{\mathrm{vol}}^{\mathrm{SDP}}$)), and illustrate Theorem 2.13(b) in more detail. Consider the ground structure and boundary conditions as shown in Figure 5 (right). The load vector consists of a single vertical force $(0, 1)$ applied at the bottom-right node. Further, let $\overline{\gamma} := 0.5$, and consider the single-load min-volume problem without vibration constraint:

$$\begin{aligned} \min_{x \in \mathbb{R}^m,\, u \in \mathbb{R}^n} & \sum_{i=1}^{m} x_i \\ \text{subject to } & K(x)u = f, \\ & f^T u \le \overline{\gamma}, \\ & x_i \ge 0, \quad i = 1, \dots, m. \end{aligned} \tag{31}$$

This problem can be formulated as a linear program [1], and thus the set $\mathcal{X}^*_{(31)}$ of solution structures of (31) is given by the set of all convex combinations of the leftmost

TABLE 1
Results of Example 4.3 for different data.

M_0	$\overline{V}$	γ^*	λ^*
0	1	1	−0.70711
0	10	0.1	−0.70711
10	1	1	−0.08761
10	10	0.1	−0.41421

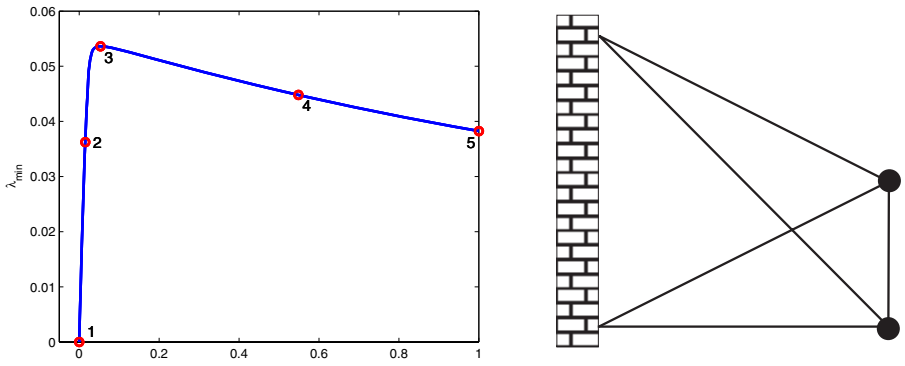


FIG. 5. *Example 4.4—graph of $\lambda_{\min}$ on interval between two structures of the same volume and compliance and ground structure (potential nodes and bars)*

and rightmost structures in Figure 6, i.e., by the set

$$\mathcal{X}^*_{(31)} = \{(1-\mu)x^{1*} + \mu x^{2*} \mid \mu \in [0,1]\},$$

where x^{1*} denotes the leftmost and x^{2*} the rightmost structures in Figure 6. We have $\mathrm{vol}(x^*) = 18$ and $c(x^*) = 1$ for all $x^* \in \mathcal{X}^*_{(31)}$. Figure 5 (left) shows the dependence of the minimum vibration eigenvalue on the parameter μ of this convex combination, i.e., a plot of the function

$$\mu \mapsto \lambda_{\min}((1-\mu)x^{1*} + \mu x^{2*})$$

over the interval $[0,1]$. The points 1–5 in the plot correspond to the structures in Figure 6, left to right. We observe that $\lambda_{\min}$ is maximized at $\mu \approx 0.0536$, i.e., at the third structure. Let us now add the vibration constraint to problem (31); thus we arrive at problem ($\mathrm{P}_{\mathrm{vol}}$). For example, put $\overline{\lambda} := 0.037$, which is the value of $\lambda_{\min}$ for the second structure in Figure 6. Then it is clear that any structure between the second and fifth trusses is a solution to problem ($\mathrm{P}_{\mathrm{vol}}$), and the vibration constraint will be inactive for the structures strictly in between. Moreover, the third truss is the structure $\hat{x}^*$ where the supremum in (19) in Theorem 2.13 is attained, i.e., the third truss is optimal for problem ($\mathrm{P}_{\mathrm{eig}}$) with the settings $\overline{V} := 18$ and $\overline{\gamma} := 1$ (according to Theorem 2.13(b)).

Example 4.5. This example shows that not only can the minimum eigenvalue function be discontinuous (see Example 2.4), but it may also behave in a non-Lipschitz way. This is slightly unexpected, given the well-known fact that the eigenvalues of the *standard* symmetric eigenvalue problem are Lipschitz.

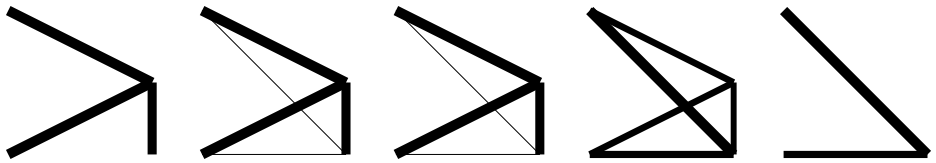


FIG. 6. *Example* 4.4—*structures corresponding to points* 1–5 *on the graph in Figure* 5.

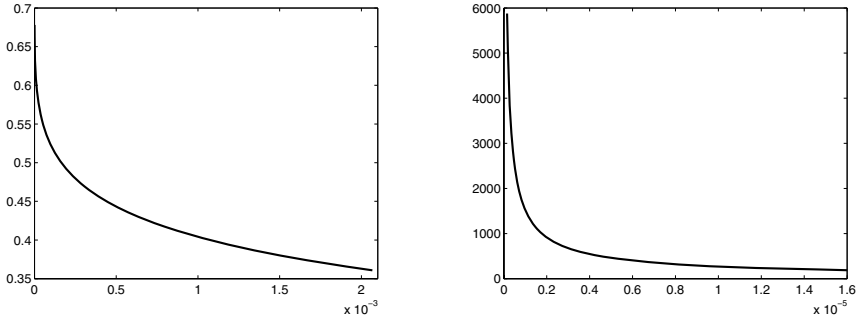


FIG. 7. *Example* 4.5 *demonstrating apparent non-Lipschitz behavior of the minimum eigenvalue function close to the boundary of the feasible region. The graph of the function (left) and its derivative (right) are shown.*

Consider again the 3×3 ground structure from Example 4.1 with all nodes connected. A horizontal force is applied at the central node. Figure 7 shows the behavior of the objective function $\lambda_{\min}(\cdot)$ of problem $(\mathrm{P}^{\mathrm{SDP}}_{\mathrm{eig}})$ with $x \geq \varepsilon > 0$; denote the solution of this problem by x_ε. The left-hand figure shows the plot of the function $\lambda_{\min}(x_\varepsilon)$ for $1.5 \cdot 10^{-7} \leq \varepsilon \leq 2 \cdot 10^{-3}$; the function looks all but Lipschitz (for smaller values of ε we were unable to compute the function value due to round-off errors). To see its behavior more clearly, we plot in the right-hand figure the derivative (computed by finite differences) in the interval $[1.5 \cdot 10^{-7}, 1.6 \cdot 10^{-5}]$; this figure confirms the non-Lipschitz behavior. When we solve the minimum eigenvalue problem $(\mathrm{P}^{\mathrm{SDP}}_{\mathrm{eig}})$ with $x \geq 0$, we obtain the optimum value $\lambda^* = -0.7071068$. Obviously, the picture is not a proof of non-Lipschitz behavior, but it is very indicative of such. The optimal trusses for $\varepsilon = 2 \cdot 10^{-3}$ and for the problem with $x \geq 0$ are shown in Figure 8 (left and right, respectively). In the first case, only bars that are not equal to the lower bound are presented. In both cases, the compliance constraint was inactive.

Example 4.6. This example demonstrates that the change in M_0 may lead to a change in the topology of the optimal structure, as has been suggested in the discussion after Proposition 2.22. We take the same ground structure, boundary conditions, and loads as in Example 4.2. Consider the minimum volume problem $(\mathrm{P}^{\mathrm{SDP}}_{\mathrm{vol}})$ with three different values of M_0, namely, 0, 10, and 100. The bounds on compliance and minimum eigenvalue are $\overline{\gamma} = 20$ and $\overline{\lambda} = 1.0 \cdot 10^{-3}$. The optimal values of V^* are, respectively, 0.05012, 0.07284, and 0.63386. In the latter case ($M_0 = 100$), the compliance constraint was inactive. The respective optimal structures are presented in Figure 9.

Example 4.7. With practical applications in mind, we also present an example of larger ground structure with multiple loads. Consider a 7×3 nodal grid with the ground structure, boundary conditions, and loads as depicted in Figure 10 (top left).

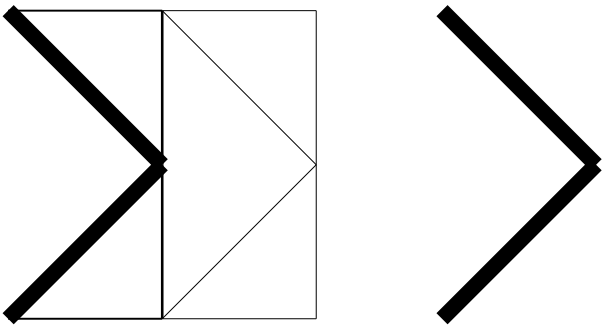


FIG. 8. *Example* 4.5—*optimal structures for* $x_i \geq 2 \cdot 10^{-3}$ *(left) and* $x_i \geq 0$ *(right).*

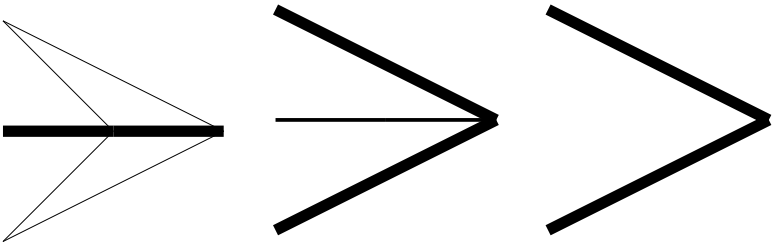


FIG. 9. *Example* 4.6 *demonstrating the dependence of the optimal structure on nonstructural mass changes; optimal results for* $M_0 = 0, 10, 100$ *are depicted left to right.*

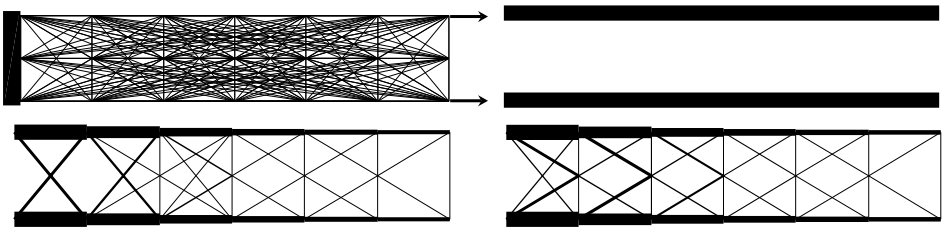


FIG. 10. *A medium-size multiple-load example (Example* 4.7*): initial layout (top left); optimal topology without (top right) and with (bottom left) vibration constraints; single-load optimal result with vibration constraints (bottom right).*

Each of the load arrows indicates an independent load case. The result of the standard minimum volume multiple-load problem (with no vibration constraints) with $\overline{\gamma} = 10$ is shown in Figure 10 (top right)—obviously resulting in two independent horizontal bars, one for each load. The volume of this structure is $V^* = 5.0$. Figure 10 (bottom left) shows the result for the multiple-load problem with a bound $\overline{\lambda} = 1.0 \cdot 10^{-3}$ on the minimum eigenvalue with the optimal volume $V^* = 7.8309$. For a comparison, we also show a result of the single-load problem (both forces considered as a single load) with $\overline{\gamma} = 20$ and $\overline{\lambda} = 1.0 \cdot 10^{-3}$; the optimal structure with $V^* = 7.6166$ is presented in Figure 10 (bottom right). All solutions were obtained by PENBMI in less than 10 seconds.

Example 4.8. We consider the same problem scenario as in Example 4.2 but with a 7×7 full ground structure with 1176 potential bars; see Figure 11 (left). Again

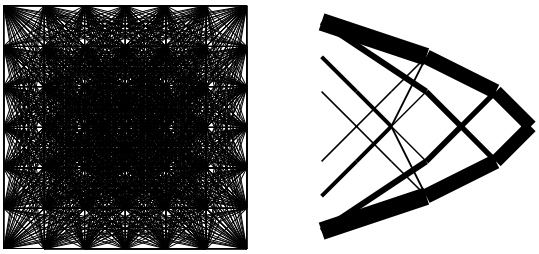


FIG. 11. *Example* 4.8—*a medium-size problem, initial layout and optimal topology.*

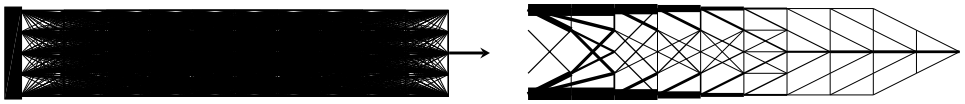


FIG. 12. *Example* 4.9—*a medium-size problem, initial layout and optimal topology.*

we solve the minimum volume problem ($\mathrm{P}_{\mathrm{vol}}^{\mathrm{SDP}}$) with $\overline{\gamma} = 1$ and $\overline{\lambda} = 5.0 \cdot 10^{-2}$ (and a nonstructural mass of size 10 at the loaded node). Figure 11 (right) shows the calculated optimal design x^*. The optimal volume is $V^* = 3.59874$, i.e., just one-half of the optimal volume of the 3×3 ground structure from before in Example 4.2. To solve the minimum volume problem by PENBMI, we needed 5 min. 16 sec. To solve the other two formulations, ($\mathrm{P}_{\mathrm{compl}}^{\mathrm{SDP}}$) and ($\mathrm{P}_{\mathrm{eig}}^{\mathrm{SDP}}$), the code needed 11 min. 41 sec. and 20 min. 15 sec., respectively. As expected, formulation ($\mathrm{P}_{\mathrm{eig}}^{\mathrm{SDP}}$) is computationally the most demanding one due to the presence of BMIs.

Example 4.9. Here we consider a medium-size example with an 11×5 ground structure, having 100 degrees of freedom and 1485 potential bars. The bounds on compliance and on the eigenvalue were $\overline{\gamma} = 20$ and $\overline{\lambda} = 5.0 \cdot 10^{-4}$. A horizontal force $(-10, 0)$ is applied at the right-middle node; see Figure 12 (left). No nonstructural mass is considered. The minimum volume problem was solved by PENBNI in 33 min. 37 sec. and resulted in the optimal structure shown in Figure 12 (right) with $V^* = 1542.65$. According to Theorem 2.8 this structure is also optimal for the minimum compliance problem ($\mathrm{P}_{\mathrm{compl}}$) with $\overline{V} := 1542.65$ and $\overline{\lambda}$ as above.

5. An extension: The multiple-mass problem. Here we propose an extension to each of the three original problem formulations, which to the best of our knowledge has not been considered before. Assume that we have n_k matrices $M_0^{(k)}$, $k = 1, \ldots, n_k$, corresponding to n_k different nonstructural masses that can be applied independently. The corresponding eigenvalue constraint extending the constraint "$\lambda_{\min}(x) \geq \overline{\lambda}$" in problem ($\mathrm{P}_{\mathrm{vol}}$) or in problem ($\mathrm{P}_{\mathrm{compl}}$) would then be stated as

$$\lambda_{\min}(x, M_0^{(k)}) \geq \overline{\lambda} \qquad \forall k = 1, \ldots, n_k,$$

where we have used the notation (28) from section 2.5 for different nonstructural mass matrices. Similarly, the objective function $\lambda_{\min}(\cdot)$ in problem ($\mathrm{P}_{\mathrm{eig}}$) becomes

$$x \mapsto \min_{1 \leq k \leq n_k} \lambda_{\min}(x, M_0^{(k)})$$

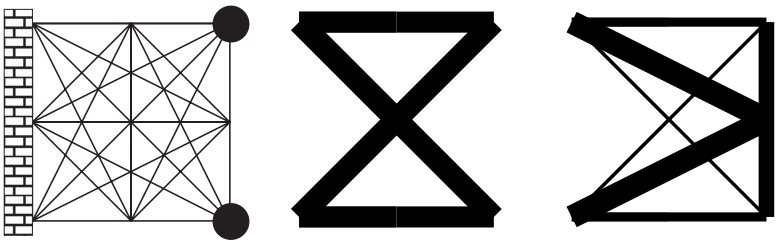


FIG. 13. *A multiple-mass problem (Example* 5.1*: initial layout (left), a "single-mass" result (middle), and a multiple-mass optimal structure (right).*

(which is to be maximized). Generalizing the SDP problem ($\mathrm{P}_{\mathrm{eig}}^{\mathrm{SDP}}$), we arrive at the following formulation possessing the same problem structure:

$$
\begin{aligned}
(32)\qquad \min_{x\in\mathbb{R}^m,\lambda\in\mathbb{R}} \; & -\lambda \\
\text{subject to } & \begin{pmatrix} \gamma & f_\ell^T \\ f_\ell & K(x) \end{pmatrix} \succeq 0, \quad \ell = 1,\dots,n_\ell, \\
& \sum_{i=1}^{m} x_i \le V, \\
& x_i \ge 0, \quad i = 1,\dots,m, \\
& K(x) - \lambda(M(x) + M_0^{(k)}) \succeq 0, \quad k = 1,\dots,n_k\,.
\end{aligned}
$$

In an analogous way we may extend the SDP problems ($\mathrm{P}_{\mathrm{vol}}^{\mathrm{SDP}}$) and ($\mathrm{P}_{\mathrm{compl}}^{\mathrm{SDP}}$) from section 3 to the case of multiple masses. Because the mathematical structure of these formulations is the same as that of problems ($\mathrm{P}_{\mathrm{vol}}^{\mathrm{SDP}}$), ($\mathrm{P}_{\mathrm{compl}}^{\mathrm{SDP}}$), and ($\mathrm{P}_{\mathrm{eig}}^{\mathrm{SDP}}$), we may use again the code PENBMI to numerically solve these problems. We finish with a numerical example.

Example 5.1. Consider a 3×3 truss with all nodes connected by potential bars. The nodes on the left-hand side are fixed in both directions, and two balls (nonstructural masses) are placed in the corners on the right-hand side; see Figure 13 (left). Figure 13 (middle) shows the optimal design for formulation ($\mathrm{P}_{\mathrm{eig}}^{\mathrm{SDP}}$) when both masses are considered a "single" nonstructural mass. Figure 13 (right) presents the result of the multiple-mass formulation (32), where the two nonstructural masses are considered being independent from each other. The volume bound in both problems was $\overline{V} := 1$, and the resulting optimal eigenvalues were $\lambda^* = 4.758 \cdot 10^{-3}$ in the single-mass case and $\lambda^* = 7.365 \cdot 10^{-3}$ in the multiple-mass case.

Appendix.

A.1. Proof of Proposition 2.3. For the proof of (a) and (b) let $x \in X$ be fixed, and let $K := K(x)$ and $M := M(x) + M_0$, for simplicity. Because M is symmetric, there exists an orthonormal basis $\{v_1,\dots,v_r\} \subset \mathbb{R}^n$ of range(M) where $r = \mathrm{rank}(M)$. Consider the matrix $P := (v_1 \cdots v_r) \in \mathbb{R}^{n\times r}$ consisting columnwise of the vectors v_j. We state the generalized eigenvalue problem

$$
P^T K P z = \lambda P^T M P z, \tag{33}
$$

with $z \in \mathbb{R}^r$.

First we show that P^TMP is positive definite. To see this, let $z \neq 0$ be arbitrary, and assume that $z^TP^TMPz = 0$. Because M is positive semidefinite, this implies $Pz = 0$. But the columns of P are linearly independent, and hence we arrive at $z = 0$, a contradiction. This shows that all eigenvalues of (33) are well defined, and (as often seen) problem (33) can be equivalently written as an ordinary eigenvalue problem

$$\widetilde{K}z = \lambda z \tag{34}$$

with $\widetilde{K} := (P^TMP)^{-1/2}P^TKP(P^TMP)^{-1/2}$.

Next we prove that λ is a well-defined eigenvalue of problem (8) if and only if it is an eigenvalue of problem (33) (and thus also an eigenvalue of $\widetilde{K}$ in (34)). First, let (λ, w) be a solution of (8) with $w \notin \ker(M)$. The latter property shows that there exist $w_1 \in \ker(M)$ and $w_2 \in \operatorname{range}(M)$, $w_2 \neq 0$, such that $w = w_1 + w_2$. Inserting $w = w_1 + w_2$ into (8) gives $Kw_1 + Kw_2 = \lambda(Mw_1 + Mw_2)$, i.e.,

$$Kw_2 = \lambda Mw_2 \tag{35}$$

due to Lemma 2.1. Notice that $w_2 \neq 0$, and thus (λ, w_2) is also a solution of (8). Because $w_2 \in \operatorname{range}(M)$, there exists $z \in \mathbb{R}^r$ such that $w_2 = Pz$. Hence, (35) becomes

$$KPz = \lambda MPz,$$

and multiplication by P^T from the left shows that (λ, z) is a solution of (33).

Conversely, let (λ, z) be a solution of (33) with $z \neq 0$. Consider $w := Pz$. Because the columns of P form a basis of $\operatorname{range}(M)$, it is $w \neq 0$ and $w \in \operatorname{range}(M)$. Through the general identity $\operatorname{range}(M)^\perp = \ker(M^T) = \ker(M)$ we see that $w \notin \ker(M)$. Moreover, as z is a solution of (33), $P^TKw = \lambda P^TMw$, which we may multiply by P from the left to end up with

$$PP^TKw = \lambda PP^TMw. \tag{36}$$

Now, Lemma 2.1 shows that $\operatorname{range}(K) \subseteq \operatorname{range}(M)$, i.e., $Kw \in \operatorname{range}(M)$. By construction, PP^T is a projection matrix onto $\operatorname{range}(M)$, and thus (36) becomes $Kw = \lambda Mw$. (Alternatively, notice that $P^TP = I_{r\times r}$. Hence, for each $\tilde{w} = P\tilde{z} \in \operatorname{range}(M)$, $PP^T\tilde{w} = PP^TP\tilde{z} = P\tilde{z} = \tilde{w}$.) As $w \notin \ker(M)$ this proves that λ is a well-defined eigenvalue of problem (8). Because $\widetilde{K} \succeq 0$, each eigenvalue λ in (34) is nonnegative, and we are done with the proof of (a).

To finish the proof of (b), we use formulation (34) and the Rayleigh quotient to see that

$$\lambda_{\min}(x) = \inf_{z\neq 0} \frac{z^T\widetilde{K}z}{z^Tz}.$$

Inserting the definition of $\widetilde{K}$, and using the substitutions $\tilde{z} := (P^TMP)^{-1/2}z$ and $w := P\tilde{z}$, we conclude

$$\lambda_{\min}(x) = \inf_{z\neq 0} \frac{z^T(P^TMP)^{-1/2}P^TKP(P^TMP)^{-1/2}z}{z^Tz} \tag{37}$$

$$= \inf_{\tilde{z}\neq 0} \frac{\tilde{z}^TP^TKP\tilde{z}}{\tilde{z}^TP^TMP\tilde{z}}$$

$$= \inf_{w\in\operatorname{range}(M):\, w\neq 0} \frac{w^TKw}{w^TMw}. \tag{38}$$

Now, for each $\tilde{u}$ with $M\tilde{u} \neq 0$ there exist $\tilde{v} \in \ker(M)$ and $\tilde{w} \in \ker(M)^\perp = \mathrm{range}(M)$ such that $\tilde{u} = \tilde{v} + \tilde{w}$. Hence, by Lemma 2.1,

$$\frac{\tilde{u}^T K \tilde{u}}{\tilde{u}^T M \tilde{u}} = \frac{\tilde{w}^T K \tilde{w}}{\tilde{w}^T M \tilde{w}}.$$

Thus we can continue (37) to (38) with

$$\lambda_{\min}(x) = \inf_{w \in \mathrm{range}(M):\, w \neq 0} \frac{w^T K w}{w^T M w} = \inf_{u:\, Mu \neq 0} \frac{u^T K u}{u^T M u},$$

which proves (b).

(c) Let us first show the "$\geq$" part. For this, take arbitrary $\lambda \in \mathbb{R}$ satisfying $K(x) - \lambda(M(x) + M_0) \succeq 0$, i.e.,

$$v^T K(x) v - \lambda v^T (M(x) + M_0) v \geq 0 \quad \forall v \neq 0, \tag{39}$$

and consider arbitrary u with $(M(x) + M_0)u \neq 0$. Then $u \neq 0$, and (39) for $v := u$ shows that

$$\frac{u^T K(x) u}{u^T (M(x) + M_0) u} \geq \lambda\,.$$

Because λ and u were arbitrary, we can write "inf" in front of the fraction and "sup" in front of λ, and the inequality remains valid, i.e.,

$$\inf_{u:\, (M(x)+M_0)u \neq 0} \frac{u^T K(x) u}{u^T (M(x) + M_0) u} \geq \sup\{\, \lambda \mid K(x) - \lambda(M(x) + M_0) \succeq 0 \,\}\,.$$

By (b) the value on the left-hand side coincides with $\lambda_{\min}(x)$, and hence we have proved "$\geq$".

The proof of the "$\leq$" part is similar: Let

$$\tilde{\lambda} := \inf_{u:\, (M(x)+M_0)u \neq 0} \frac{u^T K(x) u}{u^T (M(x) + M_0) u}.$$

Then

$$\begin{aligned}
\tilde{\lambda} &\leq \frac{u^T K(x) u}{u^T (M(x) + M_0) u} && \forall u : (M(x) + M_0)u \neq 0 \\
&\iff u^T K u - \tilde{\lambda} u^T (M(x) + M_0) u \geq 0 && \forall u : (M(x) + M_0)u \neq 0 \\
&\iff u^T K u - \tilde{\lambda} u^T (M(x) + M_0) u \geq 0 && \forall u \in \mathbb{R}^n \quad \text{(see Lemma 2.1)} \\
&\iff K(x) - \tilde{\lambda}(M(x) + M_0) \succeq 0 \\
&\iff \tilde{\lambda} \leq \sup\{\lambda \mid K(x) - \lambda(M(x) + M_0) \succeq 0\}\,.
\end{aligned}$$

(d) Let $\bar{x} \in \mathbb{R}^m$, $\bar{x} \geq 0$, and let $\{x^k\}_k$ be an arbitrary sequence such that $x^k \to \bar{x}$. We want to show that $\limsup_{x \to \bar{x}} \lambda_{\min}(x) \leq \lambda_{\min}(\bar{x})$. Take a subsequence $\{x_j^k\}_j$ of $\{x^k\}_k$ such that

$$\lim_{j \to \infty} \lambda_{\min}(x_j^k) = \bar{\lambda} := \limsup_{x \to \bar{x}} \lambda_{\min}(x)\,.$$

By definition,

$$K(x_j^k) - \lambda_{\min}(x_j^k)(M(x_j^k) + M_0) \succeq 0 \quad \forall j$$

and, passing with j to the infinity, we get

$$K(\bar{x}) - \bar{\lambda}(M(\bar{x}) + M_0) \succeq 0\,,$$

using the continuous dependence of $K(x)$ and $M(x)$ on x and closedness of the cone of positive semidefinite matrices. Hence

$$\bar{\lambda} \leq \sup\{\lambda \mid K(\bar{x} - \lambda(M(\bar{x}) + M_0) \succeq 0\} = \lambda_{\min}(\bar{x})$$

and we are done.

(e) By construction, $M(x) \succ 0$ for $x \in X_\varepsilon$. Then the pencil $(K(x), M(x) + M_0)$ is definite and we can apply general theory, saying that the eigenvalues of (8) depend continuously on parameter x [4, 29].

(f) For each $u : (M(x) + M_0)u \neq 0$, the function $u \mapsto \frac{u^T K(x) u}{u^T (M(x)+M_0) u}$ is a linear-fractional function in $(K(x), (M(x) + M_0)$, hence a quasi-linear function in variables $(K(x), (M(x) + M_0))$ (see [5]) and thus in x. Using point (b), we conclude that $-\lambda_{\min}(x)$ is quasi-convex in x, because it is the supremum of a family of quasi-linear (and thus quasi-convex) functions (here we use the fact that $-\inf g(x) = \sup -g(x)$). ∎

Remark A.1. The projection PP^T defined in the above proof takes, in fact, a particularly simple structure. Assume that $x \in X$ is given and that $\ker(M(x)) \subset \ker(M_0)$. Denote by $\mathcal{B} \subseteqq \{1, \ldots, n\}$ the degrees of freedom associated only with elements j such that $x_j = 0$ and by $\mathcal{A}$ its complement. With $k := |\mathcal{A}|$ we assume without restriction that $\mathcal{A} = \{1, \ldots, k\}$, and $\mathcal{B} = \{k + 1, \ldots, n\}$. Then $K(x)$ and $M(x) + M_0$ can be partitioned as follows:

$$K(x) = \begin{pmatrix} K_{\mathcal{A}\mathcal{A}} & K_{\mathcal{A}\mathcal{B}} \\ K_{\mathcal{B}\mathcal{A}} & K_{\mathcal{B}\mathcal{B}} \end{pmatrix}, \quad M(x) + M_0 = \begin{pmatrix} M_{\mathcal{A}\mathcal{A}} & M_{\mathcal{A}\mathcal{B}} \\ M_{\mathcal{B}\mathcal{A}} & M_{\mathcal{B}\mathcal{B}} \end{pmatrix}.$$

Clearly, $K_{\mathcal{A}\mathcal{A}} \succeq 0$; further (see Appendix A.2) $M_{\mathcal{A}\mathcal{A}} \succ 0$, and, by Lemma 2.1, $K_{\mathcal{A}\mathcal{B}} = K_{\mathcal{B}\mathcal{A}}^T = M_{\mathcal{A}\mathcal{B}} = M_{\mathcal{B}\mathcal{A}}^T = 0$ and $K_{\mathcal{B}\mathcal{B}} = M_{\mathcal{B}\mathcal{B}} = 0$ (as, e.g., $K_{\mathcal{B}\mathcal{B}} = \sum_{i:x_i=0} x_i K_i$). By this, each eigenvalue $\lambda_{\mathcal{A}}$ of the problem

$$K_{\mathcal{A}\mathcal{A}} w = \lambda_{\mathcal{A}} M_{\mathcal{A}\mathcal{A}} w$$

is a well-defined eigenvalue of problem (8).

A.2. On the representation of the mass matrix. Let $x \in X$ be given and, for simplicity of notation, assume that $M_0 = 0$ (in general, we would assume that $\ker(M(x)) \subset \ker(M_0)$). We want to show that $M(x)$, after a suitable permutation, can be partitioned into the form

$$\begin{pmatrix} M_{\mathcal{A}\mathcal{A}} & 0 \\ 0 & 0 \end{pmatrix},$$

where $M_{\mathcal{A}\mathcal{A}} \succ 0$.

LEMMA A.2. *Let $Z_i \in \mathbb{R}^{n\times n}$, $Z_i \succ 0$, and $P_i \in \mathbb{R}^{k\times n}$, $k < n$, for $i = 1, \ldots, \mu$. Then, for any $z \neq 0$,*

$$\sum_{i=1}^{\mu} P_i Z_i P_i^T z = 0 \quad \Longrightarrow \quad \sum_{i=1}^{\mu} P_i P_i^T z = 0\,.$$

Proof. From the assumption we know that $\sum_{i=1}^{\mu} z^T P_i Z_i P_i^T z = 0$ and $z^T P_i Z_i P_i^T z \geq 0$ for each i (as $P_i Z_i P_i^T \succeq 0$). Thus $z^T P_i Z_i P_i^T z = 0$ for all i and therefore

$$\|Z^{1/2} P_i^T z\|_2^2 = 0.$$

As $Z^{1/2} \succ 0$, this immediately gives $P_i^T z = 0$ for all i, and the lemma follows. ∎

Now let $\mathcal{I}$ include the indices of all nonzero components x_i of x. Without loss of generality, let us assume that the nonzero components of x are equal to one, i.e., $x_i = 1$ for $i \in \mathcal{I}$. Hence $M = \sum_{i=1}^{m} x_i P_i \widehat{M}_i P_i^T = \sum_{i \in \mathcal{I}} P_i \widehat{M}_i P_i^T$. Define the projection

$$S = I_{n \times n} - \prod_{i \in \mathcal{I}} (I_{n \times n} - P_i P_i^T)\,;$$

clearly, S projects a vector $z \in \mathbb{R}^n$ to a subspace generated by Euclidean unit vectors associated with all degrees of freedom belonging to elements $i \in \mathcal{I}$, i.e., to the space span $\{P_i P_i^T e,\ i \in \mathcal{I}\}$, where $e \in \mathbb{R}^n$ is the vector of all ones. From this definition, and from the construction of M, we immediately have that $M = SMS$. Without loss of generality, assume that S is of the form

$$S = \begin{pmatrix} I_{k \times k} & 0 \\ 0 & 0 \end{pmatrix},$$

where k is the rank of S. Hence M also has the form

$$M = \begin{pmatrix} \widetilde{M} & 0 \\ 0 & 0 \end{pmatrix},$$

with $\widetilde{M} \in \mathbb{R}^{k \times k}$.

LEMMA A.3. *$\widetilde{M}$ is positive definite.*

Proof. Assume that $Mz = 0$ for some $z \neq 0$. We need to show that $\widetilde{M}\tilde{z} = 0$ only for $\tilde{z} = 0$, where $\tilde{z}$ includes the first k components of z. By definition,

$$Mz = \sum_{i \in \mathcal{I}} P_i \widehat{M}_i P_i^T z = 0\,.$$

From the above lemma, we have that

$$\sum_{i \in \mathcal{I}} P_i P_i^T z = 0\,.$$

Now, the matrix $\sum_{i \in \mathcal{I}} P_i P_i^T$ is of the same form as S and M, and its upper-left block consists of a (full) positive diagonal. Hence $\tilde{z} = 0$. ∎

Acknowledgment. The authors would like to thank two anonymous referees for their valuable comments improving the presentation.

REFERENCES

[1] W. ACHTZIGER, A. BEN-TAL, M. BENDSØE, AND J. ZOWE, *Equivalent displacement based formulations for maximum strength truss topology design*, Impact Comput. Sci. Engrg., 4 (1992), pp. 315–345.

[2] W. ACHTZIGER, *Multiple-load truss topology and sizing optimization: Some properties of minimax compliance*, J. Optim. Theory Appl., 98 (1998), pp. 255–280.

[3] M. Bendsøe and O. Sigmund, *Topology Optimization. Theory, Methods and Applications*, Springer-Verlag, Heidelberg, 2002.
[4] R. Bhatia and R.-C. Li, *On perturbations of matrix pencils with real spectra.* II, Math. Comp., 65 (1996), pp. 637–645.
[5] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, New York, 2004.
[6] S. J. Cox and M. L. Overton, *On the optimal design of columns against buckling*, SIAM J. Math. Anal., 23 (1992), pp. 287–325.
[7] W. Dorn, R. Gomory, and M. Greenberg, *Automatic design of optimal structures*, J. de Mechanique, 3 (1964), pp. 25–52.
[8] Y. Kanno and M. Ohsaki, *Necessary and sufficient conditions for global optimality of eigenvalue optimization problems*, Struct. Multidiscip. Optim., 22 (2001), pp. 248–252.
[9] U. Kirsch, *Optimal topologies of truss structures*, Appl. Mech. Rev., 42 (1986), pp. 223–239.
[10] M. Kočvara, F. Leibfritz, M. Stingl, and D. Henrion, *A nonlinear SDP algorithm for static output feedback problems in COMPlib*, LAAS-CNRS research report 04508, LAAS, Toulouse, France, 2004.
[11] M. Kočvara and J. Outrata, *Effective reformulations of the truss topology design problem*, Optim. Engrg., 7 (2006), pp. 201–219.
[12] M. Kočvara and M. Stingl, *PENNON—a code for convex nonlinear and semidefinite programming*, Optim. Methods Softw., 18 (2003), pp. 317–333.
[13] M. Kočvara and M. Stingl, *PENBMI User's Guide. Version* 2.0, March 2005. http://www.penopt.com/.
[14] M. Kočvara, *On the modelling and solving of the truss design problem with global stability constraints*, Struct. Multidiscip. Optim., 23 (2000), pp. 189–203.
[15] A. S. Lewis and M. L. Overton, *Eigenvalue optimization*, Acta Numer., 5 (1996), pp. 149–190.
[16] J. Löfberg, *YALMIP: A toolbox for modeling and optimization in MATLAB*, in Proceedings of the CACSD Conference, Taipei, Taiwan, 2004; available from http://control.ee.ethz.ch/~joloef/yalmip.php.
[17] D. Luenberger, *Optimization by Vector Space Methods*, John Wiley, New York, 1997.
[18] E. F. Masur, *Optimal structural design under multiple eigenvalues*, Internat. J. Solids Structures, 20 (1984), pp. 211–231.
[19] M. M. Neves, H. Rodrigues, and J. M. Guedes, *Generalized topology design of structures with a buckling load criterion*, Struct. Optim., 10 (1995), pp. 71–78.
[20] M. Ohsaki, K. Fujisawa, N. Katoh, and Y. Kanno, *Semi-definite programming for topology optimization of trusses under multiple eigenvalue constraints*, Comput. Methods Appl. Mech. Engrg., 180 (1999), pp. 203–217.
[21] N. Olhoff and S. H. Rasmussen, *On single and bimodal optimum buckling loads of clamped columns*, Internat. J. Solids Structures, 13 (1977), pp. 605–614.
[22] N. Olhoff, *Optimal design with respect to structural eigenvalues*, in Theoretical and Applied Mechanics, Proceedings of the 15th International IUTAM Congress, F. P. J. Rimrott and B. Tabarott, eds., North-Holland, Amsterdam, 1980, pp. 133–149.
[23] N. L. Pedersen, *Maximization of eigenvalues using topology optimization*, Struct. Multidiscip. Optim., 20 (2000), pp. 2–11.
[24] W. Prager and J. E. Taylor, *Problems of optimal structural design*, J. Appl. Mech., 35 (1968), pp. 102–106.
[25] A. P. Seyranian, E. Lund, and N. Olhoff, *Multiple eigenvalues in structural optimization problems*, Struct. Optim., 8 (1994), pp. 207–227.
[26] A. P. Seyranian and A. A. Mailybaev, *Multiparameter Stability Theory with Mechanical Applications*, World Scientific, Singapore, 2003.
[27] G. W. Stewart, *Perturbation bounds for the definite generalized eigenvalue problem*, Linear Algebra Appl., 23 (1979), pp. 69–86.
[28] M. Stingl, *On the Solution of Nonlinear Semidefinite Programs by Augmented Lagrangian Methods*, Ph.D. thesis, Institute of Applied Mathematics, University of Erlangen–Nuremberg, Germany, 2005.
[29] T. Zhang, K. H. Law, and G. H. Golub, *On the homotopy method for perturbed symmetric generalized eigenvalue problems*, SIAM J. Sci. Comput., 19 (1998), pp. 1625–1645.

SIAM J. OPTIM.
Vol. 18, No. 4, pp. 1165–1186

SUCCESSIVE LINEAR APPROXIMATION SOLUTION OF INFINITE-HORIZON DYNAMIC STOCHASTIC PROGRAMS*

JOHN R. BIRGE† AND GONGYUN ZHAO‡

Abstract. Models for long-term planning often lead to infinite-horizon stochastic programs that offer significant challenges for computation. Finite-horizon approximations are often used in these cases, but they may also become computationally difficult. In this paper, we directly solve for value functions of infinite-horizon stochastic programs. We show that a successive linear approximation method converges to an optimal value function for the case with convex objective, linear dynamics, and feasible continuation.

Key words. stochastic programming, dynamic programming, infinite horizon, linear approximation, cutting planes

AMS subject classifications. 65K05, 90C15, 90C39, 91B28

DOI. 10.1137/060665506

1. Introduction.

Many long-term planning problems can be expressed as infinite-horizon stochastic programs. The infinite horizon often arises because of uncertainty about any specific end point (e.g., the lifetime of an individual or organization). Solving such problems with multiple decision variables and random parameters presents obvious computational difficulties. A common technique is to use a finite-horizon approximation, but even these problems become quite difficult for a practical size.

The approach in this paper is to assume stationary data and to solve for the infinite-horizon value function directly. A motivating example is an infinite-horizon portfolio problem, which involves decisions on amounts to invest in different assets and amounts to consume over time. For simple cases, such as those by Samuelson [14] for discrete time and Merton [10] for continuous time, optimality conditions can be solved directly; however, if general transaction costs and constraints on consumption or investment are present, more complex versions of the infinite-horizon problem considered here are required.

The problems in this paper also relate to the dynamic programming literature (see, for example, Bertsekas [3]) and particularly to methods for solving partially observed Markov decision processes (see the survey by Lovejoy [8]). Our method is most similar to the piecewise linear construction by Smallwood and Sondik [15] for finite horizons and the bounding approximations used by Lovejoy [9] for both finite and infinite horizons. The main differences between our model and those in [9, 15] are that we do not assume a finite action space and do not use finite-horizon approximations. Our method also does not explicitly find a policy or approximate the state space with a discrete grid; instead, we use the convexity of the value function

*Received by the editors July 19, 2006; accepted for publication (in revised form) March 29, 2007; published electronically October 10, 2007.
http://www.siam.org/journals/siopt/18-4/66550.html
†The University of Chicago Graduate School of Business, Chicago, IL 60637 (john.birge@ChicagoGSB.edu). This author's work was supported in part by the National Science Foundation under grant DMI-0200429 and by The University of Chicago Graduate School of Business.
‡Department of Mathematics, National University of Singapore, Singapore (matzgy@nus.edu.sg). This author's work was supported in part by NUS Academic Research grant R-146-000-057-112.

and the contraction properties of the dynamic programming operator in the form of an approximate value iteration.

Our approach also is similar to the linear programming approach to approximate dynamic programming (LP-ADP; see, e.g., De Farias and Van Roy [4]), but that approach again focuses on discrete state and action spaces and uses a preselected set of linear basis functions to approximate the value function. Our method uses linear support functions as an outer linearization in contrast to inner linearization in the LP-ADP approach. Our support functions are also generated within the method and achieve arbitrary accuracy. To obtain this result, our method takes advantage of continuity and convexity properties and, on all iterations, maintains bounds not available in the LP-ADP framework.

In the next section, we describe the general problem setting. Section 3 describes the algorithm and its convergence properties. Section 4 discusses the construction of the value function domain as required for algorithm convergence. Section 5 describes two examples and implementation of the algorithm. Section 6 concludes with a discussion of further issues.

2. Problem setting. We seek to find the value function V^* of the infinite-horizon problem

$$
\begin{aligned}
(2.1)\quad V^*(x) = \quad & \min_{y_1,y_2,\ldots} \quad E_{\xi_0,\xi_1,\ldots} \sum_{t=0}^{\infty} \delta^t c_t(x_t, y_t) \\
& \text{subject to (s.t.)} \ x_{t+1} = A_t x_t + B_t y_t + b_t \qquad \text{for } t = 0, 1, 2, \ldots, \\
& \qquad\qquad\qquad\quad x_0 = x.
\end{aligned}
$$

In this problem, $\xi_t = (A_t, B_t, b_t)$ is random data for stage $t = 0, 1, 2, \ldots$, A_t, B_t are matrices, and b_t is a vector. A_t is a square matrix. The equation $x_{t+1} = A_t x_t + B_t y_t + b_t$ characterizes the transition of state from stage t to $t+1$, the dynamics of the problem. $y_1, y_2, \ldots$ are controls/decision variables. State x_t and control y_t are continuous variables in certain finite-dimensional Euclidean spaces X and Y, respectively. The cost function $c_t : X \times Y \to \Re \cup \{+\infty\}$ is a generalized function. The static constraints in each stage, e.g., $(x_t, y_t) \in G_t$ for some subset G_t of $X \times Y$, are not explicitly formulated. They are implicitly captured by the effective domain of c_t, say, $dom(c_t) = G_t$. The number $0 < \delta < 1$ is a discount factor.

The above problem can be represented as

$$
\begin{aligned}
& \min_{y_0} \{c_0(x_0, y_0) + \delta E_{\xi_0} \min_{y_1} \{c_1(x_1, y_1) + \delta E_{\xi_1} \min_{y_2} \{c_2(x_2, y_2) + \ldots\}\}\} \\
& \text{s.t. } x_{t+1} = A_t x_t + B_t y_t + b_t \qquad \text{for } t = 0, 1, 2, \ldots, \\
& \qquad x_0 = x.
\end{aligned}
$$

In this paper we consider a simple version of (2.1), namely, $c_t = c$ and $(A_t, B_t, b_t) = (A, B, b)$, for all $t = 0, 1, 2, \ldots$, are identically and independently distributed random variables. For the presentation below, we assume that $\xi = (A, B, b)$ is a discrete random vector with $p_i = \text{Prob}(\xi = (A_i, B_i, b_i))$, $i = 1, \ldots, L$. (The general algorithm does not require finite realizations, but practical implementations make this assumption necessary.) The equality constraints in (2.1) characterize the dynamics of the problem. The static constraints (e.g., $y_t \geq 0$) in each stage are implicitly captured by the effective domain of the generalized function c.

The value function V^* defined by (2.1) is a solution of $V = M(V)$, where the map M (often called the *dynamic programming operator*) is defined by

$$M(V)(x) = \min_y \{c(x,y) + \delta E_\xi V(Ax + By + b)\}$$

$$= \min_y \left\{ c(x,y) + \delta \sum_{i=1}^{L} p_i V(A_i x + B_i y + b_i) \right\}. \tag{2.2}$$

Note. The problem of finding a solution of $V = M(V)$ and the infinite-horizon problem (2.1) are different. The value function V^* defined by the infinite-horizon problem is a solution to $V = M(V)$; however, the equation $V = M(V)$ may have many solutions. We will discuss this later. For the time being, we need to know only that a solution of $V = M(V)$ is equal to V^* if the effective domain of the solution coincides with $dom(V^*)$.

More precisely, let $D^* = dom(V^*)$ be compact and convex. Let $\mathcal{B}(D^*)$ be the Banach space of all functions finite on D^* and equipped with the norm $\|f\|_{D^*} = \sup_{x \in D^*}\{|f(x)|\}$.

THEOREM 2.1. *M is a contraction on $\mathcal{B}(D^*)$.*

Proof. The proof can be found, e.g., in Theorem 3.8 of [7]. □

The above theorem ensures that the equation $V = M(V)$ has a unique solution on $\mathcal{B}(D^*)$. Since V^* solves $V = M(V)$ and $V^* \in \mathcal{B}(D^*)$, V^* is the unique solution of $V = M(V)$ on $\mathcal{B}(D^*)$.

An assumption throughout this paper is that the cost function c is convex. This assumption ensures the convexity of the value function as shown below and enables the use of a cutting plane method.

LEMMA 2.2. *$M(V)$ is convex if V is convex. Consequently, V^* is convex.*

Proof. See Lemma 3.10 and Corollary 3.11 in [7]. □

In the next section, we will propose a cutting plane method to construct a piecewise linear value function V^k which approximately solves the problem $V = M(V)$. The cut generated at iteration k is a supporting plane of $M(V^k)$ at a selected point x^k. The convexity of $M(V^k)$ shown by Lemma 2.2 guarantees the existence of such cuts.

For any functions f and $g : \Re^n \to \Re \cup \{+\infty\}$, we say $f \geq g$ if $f(z) \geq g(z)$ for all $z \in \Re^n$.

3. A cutting plane method. In this section, we assume that the domain $D^* = dom(V^*)$ is known and is a compact polyhedral set. All functions are regarded as elements in $\mathcal{B}(D^*)$; thus, we will define only values of functions on D^*.

3.1. An algorithm. Our cutting plane method is based on the following observations.

LEMMA 3.1. *For any $\tilde{V}, V \in \mathcal{B}(D^*)$, $\tilde{V} \leq V$ implies $M(\tilde{V}) \leq M(V)$.*

Proof. For any $x \in D^*$,

$$\begin{aligned} M(\tilde{V})(x) &= \min_y \{c(x,y) + \delta E\tilde{V}(Ax + By + b)\} \\ &\leq \min_y \{c(x,y) + \delta EV(Ax + By + b)\} \\ &= M(V)(x). \end{aligned}$$ □

THEOREM 3.2. *Let $V \in \mathcal{B}(D^*)$, with $V \leq V^*$. Then*

(i) *$M(V) \leq V^*$;*

(ii) *$V = V^*$ if $M(V) \leq V$.*

Proof. (i) By Lemma 3.1 and $V \le V^*$, we have

$$M(V) \le M(V^*) = V^*.$$

(ii) It follows from $M(V) \le V$ and Lemma 3.1 that

$$V \ge M(V) \ge M^2(V) \ge \cdots \ge M^k(V) \ge \cdots.$$

Since M is a contraction on $\mathcal{B}(D^*)$ by Theorem 2.1, $M^k(V) \to V^*$. This shows that $V \ge V^*$. Now we have $V^* \ge V \ge V^*$, which implies $V = V^*$. □

We will employ a cutting plane algorithm to construct piecewise linear functions V^k defined by a set of cuts $\{t \ge Q^i x + q^i : i = 1, \ldots, u_k\}$ in the form

$$V^k(x) = \max\{Q^i x + q^i : i = 1, \ldots, u_k\}$$

which approximate V^* from below. Theorem 3.2 suggests adding a cut to V^k, cutting off an area where $M(V^k)(x) > V^k(x)$ so that V^{k+1} can move up towards V^*. The algorithm stops when there is no $x \in D^*$ such that $M(V^k)(x) > V^k(x)$.

ALGORITHM 3.1.

1. *Initialization: Find a piecewise linear convex function $V^0 \in \mathcal{B}(D^*)$ satisfying $V^0 \le V^*$. Set $k \leftarrow 0$.*
2. *If $V^k \ge M(V^k)$, stop; V^k is the solution. Otherwise, find a point $x^k \in D^*$ with $V^k(x^k) < M(V^k)(x^k)$.*
3. *Find a supporting hyperplane of $M(V^k)$ at x^k, say, $t = Q^{k+1}x + q^{k+1}$. Define $V^{k+1}(x) = \max\{V^k(x), Q^{k+1}x + q^{k+1}\}$.*
   $k \leftarrow k + 1$. Go to Step 2.

3.2. Details of the algorithm.

Step 1. Usually, we can find V^0 easily. For instance, if $c(x, y) \ge c_0$ for all (x, y) in its domain, then we can choose $V^0(x) = c_0/(1 - \delta)$, a constant function on D^*. It is clear that

$$V^*(x) \ge \sum_{t=0}^{\infty} \delta^t c_0 = V^0(x) \quad \forall\, x \in D^*.$$

Step 2 consists of two parts. Part 1 is the valuation of $M(V^k)(x)$, and Part 2 describes how to find a point $x^k \in D^*$ with $V^k(x^k) < M(V^k)(x^k)$.

Part 1. Assume that V^k is defined by k linear cuts, i.e., for any $x \in D^*$,

$$\begin{aligned} V^k(x) &= \max\{Q^i x + q^i : i = 1, \ldots, k\} \\ &= \min\{\theta \mid \theta \ge Q^i x + q^i,\ i = 1, \ldots, k\}. \end{aligned}$$

Then

$$\begin{aligned} M(V^k)(x) &= \min_y \left\{ c(x, y) + \delta \sum_{j=1}^{L} p_j V^k(A_j x + B_j y + b_j) \right\} \\ &= \min_{y,\theta} \left\{ c(x, y) + \delta \sum_{j=1}^{L} p_j \theta^j \mid \theta^j \ge Q^i z^j + q^i \,\forall\, i = 1, \ldots, k; \right. \\ &\qquad \left. z^j = A_j x + B_j y + b_j \in D^*,\ j = 1, \ldots, L \right\}. \end{aligned} \tag{3.1}$$

Part 2. We seek to find x^k by approximately minimizing $V^k(x) - M(V^k)(x)$ on D^*. Notice that $V^k - M(V^k)$ is a d.c. function (difference of two convex functions) on D^*. There are rich results on solving d.c. programs, originated by Horst and Tuy [6]. In Appendix A, we describe a method based on [5]. This method can find an exact global minimizer; however, finding an exact global minimizer of this d.c. function is time-consuming and thus is not recommended in general. Fortunately, for Algorithm 3.1 to work, an approximate minimizer x^k suffices (see Theorem 3.4). Selecting a computationally low-cost method for finding a sufficiently good x^k is important for enhancing the efficiency of the algorithm. The method described in Appendix A certainly is not the only choice for finding x^k. In the second example of numerical experiments, we take a small sample of 5 points and choose the best point as x^k. This simple strategy works well for the example.

Step 3. Suppose that the polytope D^* is defined by a system of inequalities $\mathbf{F}x \le \mathbf{f}$ for some matrix $\mathbf{F}$ and vector $\mathbf{f}$, i.e., $D^* = \{x \mid \mathbf{F}x \le \mathbf{f}\}$. Let

$$V^k(x) = \begin{cases} \min\{\theta \mid \mathbf{Q}^k x + \mathbf{q}^k \le \theta e\} & \text{if } \mathbf{F}x \le \mathbf{f}, \\ +\infty & \text{otherwise,} \end{cases}$$

where $\mathbf{Q}^k$ is a matrix and $\mathbf{q}^k$ and $e = (1, \ldots, 1)^T$ are vectors. $\mathbf{Q}^k x + \mathbf{q}^k$ represents the set of cuts generated up to the kth iteration.

Let (y^k, θ^k) and $\{(\lambda_j^k, \mu_j^k) : j = 1, \ldots, L\}$ be an optimal solution and optimal multipliers, respectively, of the problem

$$\min_{y,\theta} \left\{ c(x^k, y) + \delta \sum_{j=1}^{L} p_j \theta^j \mid \mathbf{Q}z^j + \mathbf{q} \le \theta^j e, \right.$$
$$\left. \mathbf{F}z^j \le \mathbf{f},\ z^j = A_j x^k + B_j y + b_j, j = 1, \ldots, L \right\}.$$

Let $\zeta^k = (\zeta_x^k; \zeta_y^k)$ be a subgradient of c at (x^k, y^k). Then

$$\xi^k = \zeta_x^k + \sum_{j=1}^{L} ((\lambda_j^k)^T \mathbf{Q} A_j + (\mu_j^k)^T \mathbf{F} A_j)$$

is a subgradient of $M(V^k)$ at x^k. (See Appendix B for the proof.)

A supporting hyperplane of $M(V^k)$ at x^k is

$$t = M(V^k)(x^k) + \xi^k (x - x^k).$$

That is,

$$Q^{k+1} = \xi^k, \qquad q^{k+1} = M(V^k)(x^k) - \xi^k x^k.$$

3.3. A modification of the algorithm. Since a global search takes considerable computational effort, a more efficient algorithm should limit global searches. We can perform a number of local search iterations between two global searches. A simple modification of the algorithm follows.

Algorithm 3.2.

1. *Initialization: Find a piecewise linear convex function* $V^0 \in \mathcal{B}(D^*)$ *satisfying* $V^0 \le V^*$. *Set* $k \leftarrow 0$.

2. *If* $V^k \geq M(V^k)$*, stop;* V^k *is the solution. Otherwise, find a point* $\bar{x} \in D^*$ *with* $V^k(\bar{x}) < M(V^k)(\bar{x})$*. Let* $V \leftarrow V^k$*.*
3. *Find a supporting hyperplane of* $M(V)$ *at* $\bar{x}$*, say,* $t = Qx+q$*. Define* $V^+(x) = \max\{V(x), Qx + q\}$*.*
4. *If the improvement of* V^+ *over* V *is larger than a certain tolerance, then let* $V \leftarrow V^+$ *and repeat Step* 3*. Otherwise, let* $V^{k+1} \leftarrow V^+$*,* $k \leftarrow k + 1$*, and go to Step* 2*.*

The time required for Step 3 is negligible compared to Step 2. The implementation on the first example in the paper shows a reduction by 3/4 of the number of iterations (of Step 2); i.e., the modified algorithm requires only 1/4 of the original number of iterations. Furthermore, it obtains better approximate solutions.

3.4. Comparison with other methods. We make comparisons in two perspectives: in improvement and in representation of an approximate value function V^k.

We refer to our method as the *successive linear approximation method* (SLAM).

Comparison between SLAM and PIM. The policy iteration method (PIM) is widely used for solving Markovian decision problems; cf. [13]. Thus we shall make a comparison between our method SLAM and PIM.

At the kth iteration, for a given V^k, PIM finds $y^k : D^* \to Y$ (which is the policy in the notation of our problem setting) by solving

$$y^k(x) = \operatorname{argmin}_y\{c(x, y) + \delta EV^k(Ax + By + b)\} \qquad \forall\, x \in D^* \tag{3.2}$$

and then finds V^{k+1} by solving

$$V^{k+1}(x) = c(x, y^k(x)) + \delta EV^{k+1}(Ax + By^k(x) + b) \quad \forall\, x \in D^*. \tag{3.3}$$

SLAM performs the kth iteration as follows: Given V^k, find a point $\bar{x}^k \in D^*$ by approximately solving

$$\max_x MV^k(x) - V^k(x) \tag{3.4}$$

and then construct V^{k+1} by adding to V^k a cutting plane at $\bar{x}^k$.

Since

$$MV^k(x) = \min_y\{c(x, y) + \delta EV^k(Ax + By + b)\},$$

finding the action $y^k(x)$ at a point x takes the same computational effort as the valuation of MV^k at x; thus, the determination of y^k, i.e., the valuation of $y^k(x)$ at all $x \in D^*$, is more expensive than the determination of $\bar{x}^k$. The construction of V^{k+1} in PIM requires solving an infinite-dimensional equation, which is again more difficult than determining a cutting plane in SLAM. In return, one can expect that each iteration of PIM improves the value function V^k more than SLAM does.

Comparison of approximation methods for the continuous-state DP. All existing approximation methods for continuous-state DP can be roughly categorized into *discrete approximations (DAs)* and *parametric approximations (PAs)*; cf. [1]. The former approximate V on a grid of D^*, and the latter approximate V by a linear combination of a set of basis functions. Our method, the *cutting plane approximation (CPA)*, approximates V by the maximum function of a set of cutting planes.

The CPA method is similar to PA in the sense that both methods use a set of continuous functions to approximate V. The superiority of PA over DA is that, in many cases, one can obtain a good global approximation to V using a small number of basis functions, whereas in high-dimensional problems the DA approach requires a very large number of grid points to obtain a comparably accurate approximation; see [1]. CPA has the same advantages as PA over DA and, moreover, is more flexible than PA, because cuts are generated where they are most needed in CPA, whereas basis functions are preselected in PA. A disadvantage of the CPA method is, however, that cuts can be accumulated rapidly. A computational challenge is to find effective ways to delete unnecessary cuts.

3.5. Convergence. If Algorithm 3.1 stops at an iteration with $V^K \geq M(V^K)$, then we have $V^K = V^*$ by Theorem 3.2.

If the algorithm does not terminate in a finite number of iterations, then the convergence of V^k to V^* is not so obvious. We notice that each cut is related to a testing point $\bar{x}$. Such a process can lead to pointwise convergence but not necessarily uniform convergence. A danger is that the limit of a pointwise convergent sequence $\{V^k\}$ may not be V^*, because V^k may be updated only in some area of the domain but not over the full domain. Our convergence analysis shall answer the following two questions: Under what conditions can the sequence $\{V^k\}$ converge uniformly? If $\{V^k\}$ converges uniformly, is the limit function equal to V^*? Our main condition for uniform convergence is that D^* is a polytope. Indeed, if D^* is an arbitrary convex compact set, one can construct a monotone increasing sequence of convex functions $\{V^k\}$ which converges pointwise but not uniformly. The assumption that D^* is a polytope ensures the continuity of the limit function of $\{V^k\}$. Then, by Theorem 7.13 in [12], $\{V^k\}$ converges uniformly; see the details in the proof of Theorem 3.4. For the second question, we will give a positive answer. The result is presented in Theorem 3.4.

LEMMA 3.3. $V^k \leq V^{k+1} \leq V^*$.

Proof. From Step 2 of Algorithm 3.1, it is obvious that $V^k \leq V^{k+1}$.

The initial function $V^0 \leq V^*$. Suppose $V^k \leq V^*$; then, by Theorem 3.2 (i), $M(V^k) \leq V^*$; thus, for any $x \in D^*$,

$$V^{k+1}(x) \leq \max\{V^k(x),\, M(V^k)(x)\} \leq V^*(x). \qquad \square$$

THEOREM 3.4. *Suppose that D^* is a polytope and suppose that, at the kth iteration, a point $x^k \in D^*$ is selected such that*

$$M(V^k)(x^k) - V^k(x^k) \geq \alpha \max\{M(V^k)(x) - V^k(x) \mid x \in D^*\}$$

for some constant $\alpha > 0$; then, $V^k \to V^*$ uniformly.

Proof. First, suppose that Algorithm 3.1 stops at a finite iteration K with $V^K \geq M(V^K)$. By Lemma 3.3, $V^K \leq V^*$. Thus, by Theorem 3.2, $V^K = V^*$.

Now suppose that Algorithm 3.1 generates a sequence $\{V^k\}$ which converges to $\tilde{V}$ pointwise.

Because $V^k \leq V^{k+1} \to \tilde{V}$, $epi(\tilde{V}) = \cap_k epi(V^k)$, which is a closed set since every V^k is closed. Thus, $\tilde{V}$ is a closed convex function on D^*. By our assumption, D^* is a polytope; thus, by Theorem 10.2 in [11], $\tilde{V}$ is continuous on D^*. By Theorem 7.13 in [12], $V^k \to \tilde{V}$ uniformly on D^*.

Assume that there exists an $\hat{x} \in D^*$ such that

$$M(\tilde{V})(\hat{x}) - \tilde{V}(\hat{x}) = 2\sigma > 0.$$

By Theorem 2.1, $M(V^k) \to M(\tilde{V})$ uniformly on D^* since $V^k \to \tilde{V}$ uniformly; thus, there exists a $\hat{k}$ such that $M(V^k)(\hat{x}) \geq M(\tilde{V})(\hat{x}) - \sigma$ for all $k \geq \hat{k}$. This yields

$$M(V^k)(\hat{x}) - V^k(\hat{x}) \geq M(\tilde{V})(\hat{x}) - \sigma - \tilde{V}(\hat{x}) = \sigma.$$

Since the supporting hyperplane at iteration k satisfies $Q^{k+1}x^k + q^{k+1} = M(V^k)(x^k)$, we have

$$\begin{aligned} V^{k+1}(x^k) - V^k(x^k) &= M(V^k)(x^k) - V^k(x^k) \\ &\geq \alpha[M(V^k)(\hat{x}) - V^k(\hat{x})] \\ &\geq \alpha\sigma. \end{aligned}$$

Thus

$$\tilde{V}(x^k) - V^k(x^k) \geq \alpha\sigma \qquad \forall\, k \geq \hat{k}.$$

The above contradicts the uniform convergence of $V^k \to \tilde{V}$ on D^*.

The contradiction implies that

$$M(\tilde{V})(x) \leq \tilde{V}(x) \qquad \forall\, x \in D^*.$$

Then, by Theorem 3.2, $\tilde{V} = V^*$. This means that V^k converges to V^* uniformly on D^*. □

4. Construction of the domain D^*. Recall that $D^* = dom(V^*)$, where V^* is the value function defined by (2.1). This section discusses how to find D^*. We do not have a complete answer for this problem; nevertheless, the method proposed can find D^* in a finite number of iterations for many cases.

4.1. An algorithm. Although the domain of a solution of $V = M(V)$ does not necessarily coincide with D^*, it does provide useful information for finding D^*. We first represent the domain of $M(V)$.

From (2.2) one can see that $x \in dom(M(V))$ if and only if there exists a y such that $c(x, y) < +\infty$ and $A_i x + B_i y + b_i \in dom(V)$ for all $i = 1, \ldots, L$. Denote

$$\begin{aligned} D_c(x) &= \{y \mid c(x, y) < +\infty\}, \\ G(x, D) &= \{y \mid A_i x + B_i y + b_i \in D\, \forall i = 1, \ldots, L\}. \end{aligned}$$

Then $x \in dom(M(V))$ if and only if $D_c(x) \cap G(x, dom(V)) \neq \emptyset$.

Denote

$$\Gamma(D) = \{x \mid D_c(x) \cap G(x, D) \neq \emptyset\}.$$

Then

$$dom(M(V)) = \Gamma(dom(V)).$$

The following are some basic properties of the operator Γ.

LEMMA 4.1.

(i) *If $D_1 \subseteq D_2$ then $\Gamma(D_1) \subseteq \Gamma(D_2)$.*

(ii) $\Gamma(D^*) = D^*$.

(iii) *If $D^* \subseteq D$, then $D^* \subseteq \Gamma(D)$.*

Proof. (i) If $D_1 \subseteq D_2$, then, for any x, $G(x, D_1) \subseteq G(x, D_2)$. Now, for any $x \in \Gamma(D_1)$, $D_c(x) \cap G(x, D_1) \neq \emptyset$. This implies $D_c(x) \cap G(x, D_2) \neq \emptyset$. Therefore, $x \in \Gamma(D_2)$.

(ii) It follows from $M(V^*) = V^*$ that $dom(M(V^*)) = dom(V^*)$, which yields $\Gamma(D^*) = D^*$ by the definition.

(iii) If $D^* \subseteq D$, then, by (i), $\Gamma(D^*) \subseteq \Gamma(D)$; hence, by (ii), we have $D^* \subseteq \Gamma(D)$. □

The following lemma shows that, if $dom(V) \subseteq dom(M(V))$, then $dom(V) \subseteq dom(V^*)$.

LEMMA 4.2. *Suppose c is bounded on its domain. If $D \subseteq \Gamma(D)$, then $D \subseteq D^*$.*

Proof. For any $x_t \in D$, we have $x_t \in \Gamma(D)$; thus,

$$\exists y_t \in D_c(x_t) : A_i x_t + B_i y_t + b_i \in D \,\forall\, i = 1, \ldots, L. \tag{4.1}$$

For any $\bar{x} \in D$, let $x_0 = \bar{x}$. There exists y_0 satisfying (4.1). For each realization of $\xi = (A, B, b)$ (where the subscript is omitted for ease of exposition), let $x_1 = Ax_0 + By_0 + b$. By (4.1), $x_1 \in D$. Since $D \subseteq \Gamma(D)$, $x_1 \in \Gamma(D)$. Therefore, there exists y_1 satisfying (4.1), and so on; so, we obtain a sequence $\{(x_t, y_t) : t = 0, 1, \ldots\}$ (here (x_t, y_t) are random vectors) satisfying $(x_t, y_t) \in dom(c)$ because $y_t \in D_c(x_t)$. Since c is bounded on its domain, $E\sum_{t=0}^{\infty} \delta^t c(x_t, y_t) < \infty$; thus, $V^*(\bar{x})$, defined by (2.1), is finite, i.e., $\bar{x} \in D^*$. Therefore, $D \subseteq D^*$. □

Suppose we use a cutting plane method to construct D^* and start with a set $D \supseteq D^*$. The above lemma suggests that one should cut off a portion of $D \setminus \Gamma(D)$ if $D \not\subseteq \Gamma(D)$.

Because

$$D^* \subseteq D_{cx} := \{x \mid D_c(x) \neq \emptyset\} = \{x \mid \exists y \text{ such that } c(x, y) < +\infty\},$$

we start with D_{cx} to find D^*. A generic cutting plane method which constructs D^* with this idea is as follows.

ALGORITHM 4.1.

1. *Let $D^0 = D_{cx}$. $k = 0$.*
2. *If $D^k \subseteq \Gamma(D^k)$, stop. Otherwise, find a cut $F^{k+1}x \leq f^{k+1}$ which cuts off a portion of $D^k \setminus \Gamma(D^k)$.*
3. *Let $D^{k+1} = D^k \cap \{x \mid F^{k+1}x \leq f^{k+1}\}$. $k \leftarrow k + 1$. Repeat.*

The algorithm stops when D^k with $D^k \subseteq \Gamma(D^k)$ is found. Question: Is $D^k = D^*$?

THEOREM 4.3. *If the algorithm terminates at a finite iteration K, then $D^K = D^*$.*

Proof. The algorithm starts with D_{cx}, which contains D^*. By Lemma 4.1, $D^* \subseteq \Gamma(D^k)$ for $k = 0, 1, \ldots, K$; thus, no cut cuts off any point of D^*. This implies $D^K \supseteq D^*$. When the algorithm stops with $D^K \subseteq \Gamma(D^K)$, Lemma 4.2 yields $D^K \subseteq D^*$; thus, $D^K = D^*$. □

Unfortunately, if Algorithm 4.1 does not stop in a finite number of iterations, the sequence $\{D^k\}$ need not converge to D^* (see Remark (iii) after Example 1 below). How to modify the algorithm to guarantee convergence is still an open question.

4.2. Generating cuts. Now we discuss Step 2 of Algorithm 4.1 in detail. First, we propose a method which finds a point $\bar{x} \in D^k \setminus \Gamma(D^k)$ and generates a cut which cuts off a portion of $D^k \setminus \Gamma(D^k)$.

For simplicity we consider only the case that $dom(c)$ is a polyhedral set. More precisely, let $dom(c) = \{(x, y) : Tx + Wy \leq r\}$. Let $D^k = \{x : F^i x \leq f^i : i = 1, \ldots, k\}$.

$\bar{x} \notin \Gamma(D^k)$ if and only if

$$\begin{aligned} T\bar{x} + Wy &\le r, \\ F^i(A_j\bar{x} + B_j y + b_j) &\le f^i, \quad i = 1, \dots, k \quad j = 1, \dots, L, \end{aligned}$$

has no feasible solution; then, by Farkas' theorem, if and only if

$$\begin{aligned} \sum_{i=1}^{k}\sum_{j=1}^{L} \pi_{ij} F^i B_j + \lambda^T W &= 0, \\ \sum_{i=1}^{k}\sum_{j=1}^{L} \pi_{ij}[F^i(A_j\bar{x} + b_j) - f^i] + \lambda^T(T\bar{x} - r) &> 0, \\ \pi \ge 0, \lambda &\ge 0, \end{aligned} \tag{4.2}$$

it has a solution.

Thus, finding a point $\bar{x} \in D^k$ such that $\bar{x} \notin \Gamma(D^k)$ is equivalent to finding a triple $(\bar{x}, \lambda, \pi)$ satisfying $\bar{x} \in D^k$ and (4.2). (Note that finding a solution to (4.2) is equivalent to determining the sign of the supremum of an indefinite quadratic function subject to linear constraints, which would also require global optimization methods as in Horst, Pardalos, and Thoai [5].)

Once a solution $(\bar{x}, \lambda, \pi)$ is found, one can construct a feasibility cut:

$$\sum_{i=1}^{k}\sum_{j=1}^{L} \pi_{ij}[F^i(A_j x + b_j) - f^i] + \lambda^T(Tx - r) \le 0, \tag{4.3}$$

i.e.,

$$F^{k+1} = \sum_{i=1}^{k}\sum_{j=1}^{L} \pi_{ij} F^i A_j + \lambda^T T, \quad f^{k+1} = \sum_{i=1}^{k}\sum_{j=1}^{L} \pi_{ij}[f^i - F^i b_j] + \lambda^T r.$$

We can add an objective to find the "best" cut in the sense, e.g., that F^{k+1} is the normal direction of a facet of D^* or, perhaps more realistically, a facet of $\Gamma(D^k)$.

Let us look at a simple example to see how $\Gamma^k(D_{cx})$ approximates D^*.

Example 1. Suppose

$$\begin{aligned} dom(c) &= \{(x, y) : x \in [-1, 1]^2,\ y \in [-\beta, \beta]\}, \\ Ax + By + b &= \alpha x + e_1 y \quad \text{for } x \in \Re^2, y \in \Re^1 \qquad \text{(deterministic)}. \end{aligned}$$

Here $0 < \alpha \in \Re^1$ and $e_1 = (1, 0)^T$, meaning that $A = \alpha I \in \Re^{2\times 2}, B = e_1 \in \Re^{2\times 1}$, and $b = (0, 0)^T$. For this example, we have

$$\begin{aligned} D_c(x) = \{y : (x, y) \in dom(c)\} &= \begin{cases} [-\beta, \beta] & \text{if } x \in [-1, 1]^2, \\ \emptyset & \text{otherwise.} \end{cases} \\ D_{cx} = \{x : D_c(x) \ne \emptyset\} &= [-1, 1]^2, \\ G(x, D) = \{y : \alpha x + e_1 y \in D\} &\qquad \text{for } D \subset \Re^2. \end{aligned}$$

Thus,

$$\begin{aligned} D_c(x) \cap G(x, D) \ne \emptyset &\iff x \in [-1, 1]^2,\ [-\beta, \beta] \cap \{y : \alpha x + e_1 y \in D\} \ne \emptyset \\ &\iff x \in [-1, 1]^2,\ (\alpha x + e_1[-\beta, \beta]) \cap D \ne \emptyset, \end{aligned} \tag{4.4}$$

where

$$\begin{aligned}\alpha x + e_1[-\beta,\beta] &= \{\alpha x + y e_1 : y \in [-\beta,\beta]\} \\ &= \left\{\alpha \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} y \\ 0 \end{pmatrix} : y \in [-\beta,\beta]\right\}.\end{aligned}$$

For any $D = \{|x_1| \le p, |x_2| \le q\}$, with $p, q \ge 0$,

$$\begin{aligned}\Gamma(D) &= \{x \in [-1,1]^2 : (\alpha x + e_1[-\beta,\beta]) \cap D \ne \emptyset\} \\ &= \{x \in [-1,1]^2 : [\alpha x_1 - \beta, \alpha x_1 + \beta] \cap [-p,p] \ne \emptyset,\ |\alpha x_2| \le q\} \\ &= \{x \in [-1,1]^2 : |x_1| \le (p+\beta)/\alpha,\ |x_2| \le q/\alpha\} \\ &= \{|x_1| \le \min\{1,(p+\beta)/\alpha\},\ |x_2| \le \min\{1, q/\alpha\}\}.\end{aligned}$$

For $0 < \alpha \le 1$,

$$\Gamma(D_{cx}) = [-1,1]^2 = D_{cx}.$$

Thus, by Lemma 4.2, $D_{cx} \subseteq D^*$, but $D^* \subseteq D_{cx}$. Thus $D^* = D_{cx} = [-1,1]^2$.

Note. For the case $\alpha = 1$, any subset D of $[-1,1]^2$ of the form $\{|x_1| \le 1,\ |x_2| \le t\}$ for some $0 \le t \le 1$ satisfies $D = \Gamma(D)$. Thus, M is a contraction on the Banach space $\mathcal{B}(D)$, and $V = M(V)$ has a solution (fixed point) V_D in $\mathcal{B}(D)$. This shows that the equation $V = M(V)$ may have many solutions.

For $1 < \alpha \le 1 + \beta$,

$$\begin{aligned}\Gamma(D_{cx}) &= \{|x_1| \le 1,\ |x_2| \le 1/\alpha\}, \\ \Gamma^2(D_{cx}) &= \Gamma(\Gamma(D_{cx})) = \{|x_1| \le 1,\ |x_2| \le 1/\alpha^2\}, \\ &\vdots \\ \Gamma^k(D_{cx}) &= \{|x_1| \le 1,\ |x_2| \le 1/\alpha^k\};\end{aligned}$$

thus, $D^* = \lim_{k\to\infty} \Gamma^k(D_{cx}) = \{|x_1| \le 1,\ x_2 = 0\} \ne D_{cx}$.

For $\alpha > 1 + \beta$,

$$\begin{aligned}\Gamma(D_{cx}) &= \left\{|x_1| \le \frac{1+\beta}{\alpha},\ |x_2| \le 1/\alpha\right\}, \\ \Gamma^2(D_{cx}) &= \left\{|x_1| \le \frac{1+\beta+\alpha\beta}{\alpha^2},\ |x_2| \le 1/\alpha^2\right\}, \\ &\vdots \\ \Gamma^k(D_{cx}) &= \left\{|x_1| \le \frac{1+\beta+\alpha\beta+\ldots+\alpha^{k-1}\beta}{\alpha^k},\ |x_2| \le 1/\alpha^k\right\};\end{aligned}$$

thus, $D^* = \lim_{k\to\infty} \Gamma^k(D_{cx}) = \{|x_1| \le \frac{\beta}{\alpha-1},\ x_2 = 0\} \ne D_{cx}$. □

Remarks.

(i) In some case ($\alpha \le 1$), $D^* = D_{cx}$, which can be obtained directly.

(ii) In some case ($\alpha > 1$), infinitely many iterations are required to reach D^*.

(iii) Suppose the cutting plane algorithm generates only single-side cuts in the case of $\alpha > 1$, e.g., only the cuts $x_2 \ge -1/\alpha^k$; then $D^k \to \{|x_1| \le 1, 0 \le x_2 \le 1\} \ne D^*$.

(iv) If we can directly generate the cut $x_2 \ge 0$ instead of infinitely many cuts $\{x_2 \ge -1/\alpha^k : k = 1,2,\ldots\}$, then we can construct D^* in 2 iterations for the problem with $1 < \alpha \le 1 + \beta$ and 4 iterations for the problem with $\alpha > 1 + \beta$.

4.3. Generating the deepest cut. A cut generated by (4.3) cuts off only points in $D^k \setminus \Gamma(D^k)$; i.e., it does not cut off any point in $\Gamma(D^k)$. As shown in Example 1, the cutting plane method using such cuts may fail to approximate D^* even with infinitely many iterations. In order to fulfill the goal in Remark (iv), we must find a deeper cut (a smaller f^{k+1}), which cuts into $\Gamma(D^k)$, hopefully reaching the boundary of D^*.

Given a D, suppose that we have obtained a cut $d^T x \le t_0$ which cuts off a portion of $D \setminus \Gamma(D)$ (but does not cut into $\Gamma(D)$). Denote $D_t := D \cap \{d^T x \le t\}$. If the plane $d^T x = t_0$ has touched the boundary of $\Gamma(D)$, then no point of $D_{t_0} \setminus \Gamma(D)$ can be cut off by any cut of the form $d^T x \le t$ for arbitrary t. However, since $\Gamma(D_{t_0})$ is smaller than $\Gamma(D)$, a portion of $D_{t_0} \setminus \Gamma(D_{t_0})$ may be cut off by some cut $d^T x \le t$. We wish to find $\bar{t} < t_0$ such that no point of $D_{\bar{t}} \setminus \Gamma(D_{\bar{t}})$ can be cut off by any cut of the form $d^T x \le t$. In other words, the plane $d^T x = \bar{t}$ touches the boundary of $\Gamma(D_{\bar{t}})$ (a supporting plane of $\Gamma(D_{\bar{t}})$). Thus, for any $t > \bar{t}$, the plane $d^T x = t$ does not touch $\Gamma(D_t)$, and, for any $t \le \bar{t}$, the half-space $d^T x \le t$ intersects with $\Gamma(D_t)$. The latter means that there exists $x \in \Gamma(D_t)$ satisfying $d^T x \ge t$. The latter interpretation suggests determining $\bar{t}$ by the following linear program:

$$\bar{t} = \max \{t \mid d^T x \ge t,\ x \in \Gamma(D_t)\}. \tag{4.5}$$

Because $d^T x \le t_0$ does not cut into $\Gamma(D)$ (then does not cut into $\Gamma(D_t)$), there exists no x satisfying $d^T x \ge t$ and $x \in \Gamma(D_t)$ if $t > t_0$. This implies that $\bar{t} \le t_0$. Therefore, the cut $d^T x \le \bar{t}$ is deeper than the cut $d^T x \le t_0$.

On the other hand, the following lemma guarantees that the cut $d^T x \le \bar{t}$ will not cut off any point in D^*.

LEMMA 4.4. *Suppose that $D^* \subset D$. Let $\bar{t}$ be the optimal objective value of problem (4.5); then, $d^T x \le \bar{t}$ is satisfied by all $x \in D^*$.*

Proof. Let

$$x^* = \operatorname{argmax}\{d^T x \mid x \in D^*\}, \qquad t^* = d^T x^*.$$

Because $x^* \in D^*$, there exist $\{(x_t, y_t) : t = 0, 1, 2, \ldots\}$ such that

$$V^*(x^*) = E\left[\sum_{t=0}^{\infty} \delta^t c(x_t, y_t)\right] < \infty,$$
$$x_{t+1} = Ax_t + By_t + b, \qquad x_0 = x^*.$$

For any integer $K \ge 0$, let $\tilde{x}_l = x_{l+K}$ and $\tilde{y}_l = y_{l+K}$. Because $E[\sum_{t=K}^{\infty} \delta^{t-K} c(x_t, y_t)] < \infty$, we have

$$V^*(x_K) \le E\left[\sum_{l=0}^{\infty} \delta^l c(\tilde{x}_l, \tilde{y}_l)\right] < \infty.$$

Therefore, $x_K \in D^*$.

Now $y_0 \in D_c(x^*)$ follows from $c(x^*, y_0) < \infty$ and $y_0 \in G(x^*, D^*)$ follows from $x_1 = A_i x^* + B_i y_0 + b_i \in D^*$ for every $i = 1, \ldots, L$. Thus $x^* \in \Gamma(D^*)$. Because $D^* \subseteq D$ and $D^* \subseteq \{x : d^T x \le t^*\}$, we have $D^* \subseteq D_{t^*}$. Therefore, $x^* \in \Gamma(D_{t^*})$. This, together with $d^T x^* = t^*$, shows that (x^*, t^*) is a feasible point of (4.5); thus $t^* \le \bar{t}$, from which the claim of the lemma follows. ∎

The following example shows the effect of the deepest cut.

Example 1 (continued). Consider $\alpha > 1$. Consider the cut of the form $-x_2 \le t$ for some $t \in R$ to be determined. So, $d = (0, -1)^T$ in (4.5). Start with $D = D_{cx} = [-1, 1]^2$; then

$$D_t = \{x \in [-1, 1]^2 : -x_2 \le t\}.$$

Linear program (4.5) is

$$\begin{aligned} \bar{t} = \max\, & t \\ \text{s.t. } & -x_2 \ge t, \\ & -1 \le \alpha x_1 + y \le 1, \\ & -t \le \alpha x_2 \le 1, \\ & -\beta \le y \le \beta. \end{aligned}$$

A feasible solution must satisfy

$$-t/\alpha \le x_2 \le -t.$$

This can be satisfied only when $t \le 0$ since $\alpha > 1$. Thus, we have $\bar{t} = 0$. The cut $-x_2 \le 0$ reaches the bottom of D^*. With one more cut from above ($d = (0, 1)^T$), D^* will be completely determined for the case of $1 < \alpha \le 1 + \beta$.

For the case of $\alpha > 1 + \beta$, suppose we have $d = (1, 0)^T$. Then

$$D_t = \{x \in [-1, 1]^2 : x_1 \le t\}.$$

Linear program (4.5) is

$$\begin{aligned} \bar{t} = \max\, & t, \\ \text{s.t. } & x_1 \ge t, \\ & -1 \le \alpha x_1 + y \le t, \\ & -1 \le \alpha x_2 \le 1, \\ & -\beta \le y \le \beta. \end{aligned}$$

Feasible solutions must satisfy

$$t \le x_1 \le \frac{t + \beta}{\alpha}.$$

This implies

$$t \le \frac{\beta}{\alpha - 1}.$$

Thus $\bar{t} = \frac{\beta}{\alpha - 1}$; so we obtain a cut $x_1 \le \frac{\beta}{\alpha - 1}$ which cuts exactly to the boundary of D^* on the right. One more cut from the left ($d = (-1, 0)^T$) will completely determine D^*; hence, in total, we need only 4 cuts. □

5. Examples.

5.1. Infinite-horizon portfolio. As noted in the introduction, this work was motivated by solving infinite-horizon investment problems that face long-enduring institutions. We will demonstrate how the algorithm performs on a small example where an infinite-horizon optimum can be found analytically (as done, for example, in [14, 10]). The goal is to maximize the discounted expected utility of consumption over an infinite horizon. The decisions in each period are how much to consume and how much to invest in a risky asset (or in a variety of assets).

The state variable x in this case corresponds to wealth or the current market value of all assets. The control variable y has two components: y_1, which corresponds to consumption, and y_2, which corresponds to the amount invested in a risky asset with random return ξ. The assumption in this model is that any remaining funds, after consuming y_1 and investing y_2 in the risky asset, are invested in a risk-free asset (e.g., U.S. Treasury bills) with a known rate of return r. For this model, $c(x, y)$ is either ∞ if $x < 0$ or $-y_1^\gamma/\gamma$ for some nonzero parameter $\gamma < 1$, giving (the negative of) the common utility function with constant relative risk aversion (i.e., such that risk preferences do not depend on the level of wealth).

With these assumptions, $M(V)$ takes the following form (for $x \geq 0$):

$$M(V)(x) = \min_y \left\{ -y_1^\gamma/\gamma + \delta \sum_{i=1}^{L} p_i V((1+r)x - (1+r)y_1 + (\xi_i - r)y_2) \right\}, \tag{5.1}$$

where ξ_i is the ith realization of the random return with probability p_i. The solution V^* of $M(V) = V$ can be found analytically by observing that the optimal value function is proportional to x^γ (by, for example, considering the limiting case of a finite-horizon problem). We then have that

$$\begin{aligned}
&\sum_{i=1}^{L} p_i V((1+r)x - (1+r)y_1 + (\xi_i - r)y_2) \\
&= -K \sum_{i=1}^{L} p_i ((1+r)x - (1+r)y_1 + (\xi_i - r)y_2)^\gamma \\
&= -K(x - y_1)^\gamma \sum_{i=1}^{L} p_i ((1+r) + (\xi_i - r)[y_2/(x - y_1)])^\gamma \\
&= -K(x - y_1)^\gamma \sum_{i=1}^{L} p_i ((1+r) + (\xi_i - r)z)^\gamma,
\end{aligned}$$

where $z = \frac{y_2}{x - y_1}$ is the fractional risky investment after consuming y_1 and K is some positive constant. The optimal z^* then must solve

$$\sum_{i=1}^{L} p_i \gamma (\xi_i - r)((1+r) + (\xi_i - r)z)^{\gamma - 1} = 0, \tag{5.2}$$

which is independent of y_1. With $\bar{V}^* = -\sum_{i=1}^{L} p_i ((1+r) + (\xi_i - r)z^*)^\gamma$, optimal y_1^* now must solve

$$-y_1^{\gamma - 1} + \delta \gamma K \bar{V}^* (x - y_1)^{\gamma - 1} = 0 \tag{5.3}$$

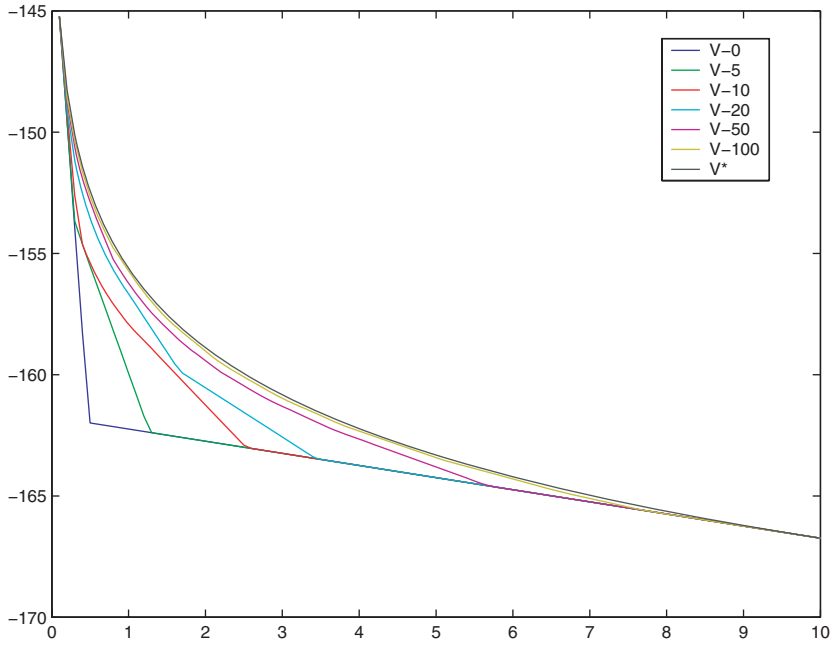


FIG. 1. *Value function approximations for portfolio example* $\delta = 1/1.25$. *The horizontal axis stands for the state variable* x *and the vertical axis for the value function* V.

or

$$y_1^* = x\frac{(\delta\gamma K\bar{V}^*)^{1/(\gamma-1)}}{1+(\delta\gamma K\bar{V}^*)^{1/(\gamma-1)}} = xw^* \tag{5.4}$$

for an optimal consumption fraction w^*. The last step is to find K from $M(V) = V$, using

$$-x^\gamma((w^*)^\gamma/\gamma + \delta K\bar{V}^*(1-w^*)^\gamma) = M(V(x)) = V(x) = -Kx^\gamma \tag{5.5}$$

to obtain $K = \frac{((\delta\bar{V}^*)^{1/(\gamma-1)}-1)^{\gamma-1}}{\delta\gamma\bar{V}^*}$.

While this function can be found explicitly (up to solving the nonlinear equation (5.2)), various other constraints on investment (such as transaction costs and limits on consumption changes from period to period) make analytical solutions impossible. We use the analytical solution here to observe Algorithm 3.1's performance and convergence behavior. We also use the analytical solution to derive initial upper bounding approximations. For our test, we use the linear supports of V^* at $x = 0.1$ and $x = 10$ as initial cuts and restrict our search in x to the interval $[0.1, 10]$, although the feasible region is unbounded.

For our test, we used $\gamma = 0.03$, $r = 0.05$, and ξ_i chosen as a discrete approximation of the lognormal return distribution with a mean return of 0.08 and a standard deviation of 0.4. Algorithm 3.1 was implemented in MATLAB using `fmincon` to solve the optimization subproblems and a linesearch to find x^k in Step 2. We tried different values for the discount factor δ. The results for $\delta = \frac{1}{1.25}$ appear in Figure 1, which includes V^0, V^5, V^{10}, V^{20}, V^{50}, V^{100}, and V^*. In this case, after 100 iterations, the approximation almost perfectly matches the true infinite-horizon value function.

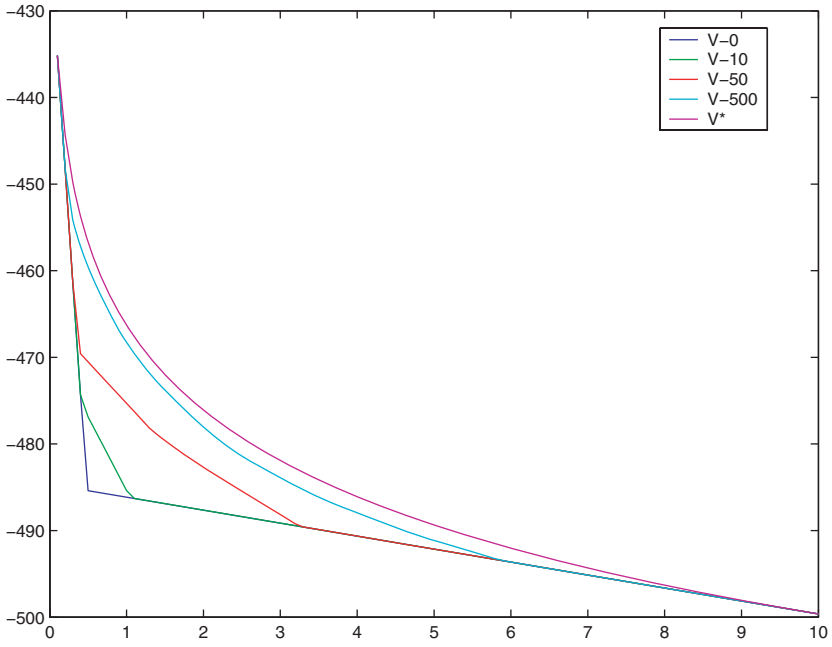


FIG. 2. *Value function approximations for portfolio example* $\delta = 1/1.07$. *The horizontal axis stands for the state variable* x *and the vertical axis for the value function* V.

With a larger δ, the value of K increases rapidly as $(\delta \bar{V}^*)^{1/(\gamma-1)}$ approaches one. For $\delta = \frac{1}{1.25}$, K is 155.6, while for $\delta = \frac{1}{1.07}$, K is 466.3. The result is that larger δ values (corresponding to lower discount rates) require additional iterations of Algorithm 3.1 to approach V^*. The results for the same data as in Figure 1 except with $\delta = \frac{1}{1.07}$ appear in Figure 2. After 500 iterations, the approximating V^{500} agrees relatively well with V^* as shown in the figure but has not converged to nearly the same accuracy as the approximations (with fewer iterations) in Figure 1.

For this example, Algorithm 3.2 gains a notable speedup. For the instance with $\delta = \frac{1}{1.25}$, it requires only 25 iterations to obtain the approximating V^{25} which is as good as V^{100} obtained by Algorithm 3.1. For the case of $\delta = \frac{1}{1.07}$, Algorithm 3.2 requires only about 1/4 of the iterations needed by Algorithm 3.1. Moreover, Algorithm 3.2 can generate more accurate approximating value functions.

Understanding the numerical behavior of Algorithm 3.1 and finding mechanisms to speed convergence should be topics for further investigation. Comparing V^k to V^* in the figures shows how the algorithm forces a closer approach in some areas of the curve over others. The behavior of the algorithm is generally to move along the curve V^k to create tighter cuts and then to repeat that process with less improvement on a new sequence of iterations. These observations suggest that procedures with multiple cut generation and tightening tolerances should be considered for accelerating convergence.

5.2. Quadratic-linear stochastic control problems. This example is intended to test the algorithm for high-dimensional problems. The quadratic-linear stochastic control problems are well known. We will use these examples following section 6.5 of [2]. The value function in a quadratic-linear stochastic control problem

is defined as follows:

$$V^*(x) = \min_{y_1,y_2,\ldots} E_{\xi_0,\xi_1,\ldots} \sum_{t=0}^{\infty} \delta^t [x_t^T Q x_t + y_t^T R y_t]$$
$$\text{s.t.} \quad x_{t+1} = Ax_t + By_t + b \qquad \text{for } t = 0, 1, 2, \ldots,$$
$$x_0 = x.$$

In the above expression, $x_t \in \Re^n$, $y_t \in \Re^m$, and the matrices A, B, Q, R are given and have appropriate dimensions. We assume that Q is a symmetric positive semidefinite matrix and that R is also symmetric and positive definite. The disturbances $\xi_t = b$ form independent random vectors with given probability distributions that do not depend on x_t, y_t. Furthermore, the vector b has zero mean and finite second moment. The controls y_t are unconstrained. The problem above represents a popular formulation of a regulation problem whereby we desire to keep the state of the system close to the origin. Such problems are common in the theory of automatic control of a motion or a process. For computational purposes, we assume $\xi = b$ is a discrete random vector with $p_i = \text{Prob}(\xi = b_i)$, $i = 1, \ldots, L$.

The value function V^* is the solution of $M(V) = V$, where the mapping M is determined by

$$M(V)(x) = \min_y \left\{ x^T Q x + y^T R y + \delta \sum_{i=1}^{L} p_i V(Ax + By + b_i) \right\}.$$

We choose the quadratic-linear stochastic control problem because its exact solution on $\Re^n$ can be computed. In the following, to be consistent with standard notation for these problems, we reuse the notation K and c as a matrix and a constant, respectively. To find the exact solution for V^*, we first compute a symmetric positive semidefinite matrix K by solving the algebraic matrix Riccati equation:

$$K = A^T[\delta K - \delta^2 K B(\delta B^T K B + R)^{-1} B^T K]A + Q.$$

Then the exact value function V^* is given by

$$V^*(x) = x^T K x + c,$$

where

$$c = \frac{\delta}{1-\delta} \sum_{i=1}^{L} p_i b_i^T K b_i.$$

The above formula determines the exact value function on $\Re^n$; however, a numerical method can determine only a value function on a bounded set. In this example, we will find the value function on the box $D^* = [-t, t]^n$. The solutions of $M(V) = V$ on $\mathcal{B}(\Re^n)$ and $\mathcal{B}([-t, t]^n)$ need not be the same; however, for reasonably large t, the two solutions should be close. Thus, we still use the exact value function on $\mathcal{B}(\Re^n)$ as the reference for numerically generated value functions.

Finding a point $\bar{x}$ which maximizes $M(V)(x) - V(x)$ is computationally very expensive. The method for minimizing a d.c. function described in Appendix A can be applied to finding $\bar{x}$; however, this is difficult to program and expensive in computation. Thus, we use the following approach. We randomly sample J points,

x^j, $j = 1, \ldots, J$ (e.g., $J = 5$). At each point x^j, we evaluate $M(V)(x^j)$ and compute the gradient of $M(V)$ (or a subgradient if $M(V)$ is not smooth) ξ^j, which generates at the point a supporting plane H^j of $M(V)$. Then we find a point $\bar{x}^j$ by maximizing $H^j(x) - V(x)$, which can be viewed as a local approximation of $M(V)(x) - V(x)$. Suppose that the current approximate value function

$$V(x) = \max\{Q^i x + q^i : i = 1, \ldots, p\}, \qquad x \in [-t, t]^n.$$

Using the supporting plane formula from section 3.2, we have

$$H^j(x) = M(V)(x^j) + \xi^j(x - x^j).$$

Thus, maximizing $H^j(x) - V(x)$ can be formulated as a linear program

$$\sigma^j = \max\{M(V)(x^j) + \xi^j(x - x^j) - \theta : x \in [-t, t]^n,\ Q^i x + q^i \le \theta,\ i = 1, \ldots, p\};$$

then $\bar{x}$ is chosen as the point among $\bar{x}^j$, $j = 1, \ldots, J$, with the largest σ^j.

We implemented SLAM for the quadratic-linear stochastic control problem with various dimensions and summarize the results in the tables below. The accuracy of an approximate solution to the exact solution is measured by the *relative error*,

$$\|V^* - V\|_{D^*} / \|V^*\|_{D^*}.$$

Although the exact value function $V^*(x) = x^T K x + c$ is known, computing the approximation error is still difficult due to the high dimension of $D^* = [-t, t]^n$. To estimate the error, we generate a sample of 5000 points plus points on the line of the eigenvector corresponding to the maximum eigenvalue of K and estimate $\|f\|_{D^*}$ by $\max\{|f(x)| : \text{for all sample points } x\}$. This sample size is generally small for high-dimensional problems, causing potential inaccuracy in the errors shown in the table. The result can, however, be understood as a tendency. For each dimension, we computed 3 instances. The errors shown in the table are the average of the 3 instances.

$(n, m) =$	Number of Iterations: 10	25	50	100	200
$(1, 1)$	0.3540	0.1238	0.0473	0.0125	0.0037
$(2, 2)$	0.1680	0.0750	0.0466	0.0191	0.0105
$(4, 2)$	0.1444	0.0649	0.0541	0.0309	0.0196
$(4, 4)$	0.1140	0.0512	0.0380	0.0190	0.0153
$(10, 5)$	0.1311	0.1255	0.0785	0.0499	0.0455
$(10, 10)$	0.1429	0.1403	0.0868	0.0751	0.0670

Relative Errors

The experiments show that fairly good solutions can be obtained for low-dimensional instances. For the high-dimensional cases, errors are rapidly reduced in the first several iterations but are not clearly reduced after the initial iterations through the 200th iteration. We display only the first iterations for the examples of dimensions $(n, m) = (20, 20)$ and $(50, 50)$.

# Iterations	1	2	3	4	6	8	200
(20,20)	1.0000	0.3935	0.3926	0.1383	0.1221	0.1159	0.1126
(50,50)	1.0000	0.7774	0.3442	0.1794	0.0939	0.0925	0.0925

Relative Errors

A positive aspect of these tests is that the method appears to work well for high-dimensional problems in the sense that it can find a relatively good solution in the first few iterations, with relative errors reduced to around 0.1. Considering that it is not possible for a direct state-discretization method to work for a 20-dimensional problem (since that requires q^{20} points if each dimension is discretized into q points), our method can be considered as a useful alternative for solving high-dimensional stochastic dynamic programs for which the curse of dimensionality is a major concern.

The negative result from this example is that, after a few iterations, the method converges very slowly, particularly for high-dimensional problems. This can be seen from the examples of dimensions $(n, m) = (20, 20)$ and $(50, 50)$. Useful cuts continue to be added in these examples, but, while reducing local errors about the iteration points, errors still remain unaffected in other portions of the domain.

6. Observations and future issues. We have described an algorithm for solving a general class of discrete-time convex infinite-horizon optimization problems and demonstrated the method on two simple examples. As the first example demonstrates, many iterations may be required to achieve accuracy for highly nonlinear value functions. The second example, however, demonstrates that approximate solutions converge very quickly at the beginning and can produce approximate solutions with a relative error of around 0.1, even for high-dimensional problems.

Since each iteration involves additional optimization steps, the selection of $\bar{x}$ (or potentially multiple points on each iteration) has a critical effect on performance. We mentioned the d.c. methods as one possibility for finding good points, but other methods that require fewer function evaluations may also be useful. In the second example, we select $\bar{x}$ from a small sample. The results from this example are encouraging. To see if this selection is effective in general, more numerical experiments are needed. The effect of different options also requires further study.

Our method relies on identification of the feasible domain D^*. In that case, we are left with the following questions:

(i) Under what condition is D^* a polytope?
(ii) Can Algorithm 4.1 terminate in a finite number of iterations if D^* is a polytope?
(iii) How can one modify the algorithm if D^* is not a polytope?

These questions and further implementation issues are additional subjects for future research.

Appendix A. This appendix describes a method for finding a minimizer to the problem

$$\min_{x \in D^*} V(x) - M(V)(x),$$

where both V and $M(V)$ are convex. This problem is equivalent to

$$\begin{aligned} &\min x_{n+1}, \\ x, x_{n+1} &: V(x) - M(V)(x) - x_{n+1} \le 0, \\ &x \in D^*, \end{aligned}$$

which is equivalent to

$$\begin{aligned} &\min x_{n+1}, \\ x, x_{n+1}, x_{n+2} &: V(x) - x_{n+1} - x_{n+2} \le 0, \\ &M(V)(x) - x_{n+2} \ge 0, \\ &x \in D^*. \end{aligned}$$

Suppose that

$$V(x) = \max\{Q^i x + q^i : i = 1, \ldots, k\} \qquad \forall\, x \in D^* \subset \Re^n;$$

then the above problem is equivalent to

$$\begin{aligned} &\min x_{n+1}, \\ x, x_{n+1}, x_{n+2} : &Q^i x + q^i - x_{n+1} - x_{n+2} \le 0, \qquad i = 1, \ldots, k, \\ &x \in D^*, \\ &M(V)(x) - x_{n+2} \ge 0. \end{aligned}$$

The first $k+1$ constraints define a polyhedral set, denoted by D. (In order to use the algorithm described in [5], D should be bounded. This can be done by adding appropriate lower and upper bounds on x_{n+1} and x_{n+2}, without changing the solution of the minimization problem.) The function in the $k+2$nd constraint is convex, denoted by g. Such a program can be solved by Algorithm 4.1 in [5]. Here we describe it briefly. The algorithm solves

$$\begin{aligned} &\min c^T x \\ &\text{s.t. } x \in D,\ g(x) \ge 0. \end{aligned} \tag{A.1}$$

Initialization: Solve $\min\{c^T x : x \in D\}$ to obtain $x^0 \in D$. Assume $g(x^0) < 0$ (otherwise, x^0 is an optimal solution to (A.1)). Construct a simple polytope S_0, e.g., a simplex, containing D, such that x^0 is a vertex of S_0; compute the vertex set $V(S_0)$ of S_0. $j \leftarrow 0$.

Iterations: Choose $z \in V(S_j)$ satisfying $g(z) = \max\{g(x) : x \in V(S_j)\}$.

If $g(z) = 0$ and $z \in D$, then stop; z is an optimal solution.

Otherwise, apply the simplex algorithm with starting vertex z to solve $\min\{c^T x : x \in S_j\}$ until an edge $[u, v]$ of S_j is found such that $g(u) \ge 0$ and $g(v) < 0$ and $c^T v < c^T u$; compute the intersection point s of $[u, v]$ with $\{x : g(x) = 0\}$.

If $s \in D$, then $S_{j+1} \leftarrow S_j \cap \{x : c^T x \le c^T s\}$.

If $s \notin D$, then $S_{j+1} \leftarrow S_j \cap \{x : l(x) \le 0\}$, where $l(x) \le 0$ is one of the linear constraints defining D satisfying $l(s) > 0$.

$j \leftarrow j + 1$; repeat.

Appendix B. Define the function $\rho = M(V)$ by

$$\rho(x) = \min_{y,\theta} \left\{ c(x, y) + \delta \sum_{j=1}^{L} p_j \theta^j \mid \mathbf{Q} z^j + \mathbf{q} \le \theta^j e,\ \mathbf{F} z^j \le \mathbf{f}, \right.$$
$$\left. z^j = A_j x + B_j y + b_j,\ j = 1, \ldots L \right\},$$

and let $(\bar{y}, \bar{\theta})$ and $(\bar{\lambda}, \bar{\mu})$ be an optimal solution and multiplier, respectively, of the above minimization problem. We will show that

$$\xi = \zeta_x + \sum_{j=1}^{L} (\bar{\lambda}_j^T \mathbf{Q} A_j + \bar{\mu}_j^T \mathbf{F} A_j)$$

is a subgradient of ρ at $\bar{x}$, where $\zeta = (\zeta_x, \zeta_y)$ is a subgradient of c at $(\bar{x}, \bar{y})$.

We can write

$$\begin{aligned}\rho(x) = \min\ & c(x,y) + \delta \sum_{j=1}^{L} p_j \theta^j, \\ & y, \theta : \mathbf{Q}(A_j x + B_j y + b_j) + \mathbf{q} \le \theta^j e, \\ & \qquad \mathbf{F}(A_j x + B_j y + b_j) \le \mathbf{f}, \qquad j = 1, \dots, L, \\ = \max\ & h(\lambda, \mu; x) \\ \text{s.t. } & \lambda_j^T e = \delta p_j, \qquad j = 1, \dots, L, \\ & \lambda \ge 0,\ \mu \ge 0, \end{aligned} \tag{B.1}$$

where $\lambda = (\lambda_1, \dots, \lambda_L)$, $\mu = (\mu_1, \dots, \mu_L)$, and

$$\begin{aligned} h(\lambda, \mu; x) = \min_y\ & c(x,y) + \sum_{j=1}^{L} [\lambda_j^T (\mathbf{Q}(A_j x + B_j y + b_j) + \mathbf{q}) \\ & + \mu_j^T (\mathbf{F}(A_j x + B_j y + b_j) - \mathbf{f})]. \end{aligned} \tag{B.2}$$

Let $(\bar\lambda, \bar\mu)$ be the optimal solution of the problem (B.1) with $x = \bar x$. Then $\rho(\bar x) = h(\bar\lambda, \bar\mu; \bar x)$.

The necessary and sufficient condition for $\bar y$ to be an optimal solution of the problem (B.2) (given $(\bar\lambda, \bar\mu; \bar x)$) is that there exists a subgradient $\zeta = (\zeta_x, \zeta_y)$ of c at $(\bar x, \bar y)$ such that

$$\zeta_y + \sum_{j=1}^{L} [\bar\lambda_j^T \mathbf{Q} B_j + \bar\mu_j^T \mathbf{F} B_j] = 0. \tag{B.3}$$

For fixed $(\lambda, \mu) = (\bar\lambda, \bar\mu)$ and for any x, denote by y_x an optimal solution of (B.2). Then

$$\begin{aligned} \rho(x) &= \max_{\lambda,\mu} \{h(\lambda, \mu; x) \mid \lambda_j^T e = \delta p_j,\ j = 1, \dots, L;\ \lambda \ge 0,\ \mu \ge 0\} \\ &\ge h(\bar\lambda, \bar\mu; x) \\ &= c(x, y_x) + \sum_{j=1}^{L} [\bar\lambda_j^T (\mathbf{Q}(A_j x + B_j y_x + b_j) + \mathbf{q}) + \bar\mu_j^T (\mathbf{F}(A_j x + B_j y_x + b_j) - \mathbf{f})]. \end{aligned}$$

Because c is convex,

$$c(x,y) \ge c(\bar x, \bar y) + \zeta_x (x - \bar x) + \zeta_y (y - \bar y).$$

Note that $y_{\bar x} = \bar y$ and

$$\begin{aligned} \rho(\bar x) &= h(\bar\lambda, \bar\mu; \bar x) \\ &= c(\bar x, \bar y) + \sum_{j=1}^{L} [\bar\lambda_j^T (\mathbf{Q}(A_j \bar x + B_j \bar y + b_j) + \mathbf{q}) + \bar\mu_j^T (\mathbf{F}(A_j \bar x + B_j \bar y + b_j) - \mathbf{f})]. \end{aligned}$$

Thus,

$$\begin{aligned}\rho(x) &\geq \rho(\bar{x}) + \zeta_x(x-\bar{x}) + \zeta_y(y_x - \bar{y}) \\ &\quad + \sum_{j=1}^{L}[\bar{\lambda}_j^T(\mathbf{Q}A_j(x-\bar{x}) + \mathbf{Q}B_j(y_x-\bar{y})) + \bar{\mu}_j^T(\mathbf{F}A_j(x-\bar{x}) + \mathbf{F}B_j(y_x-\bar{y}))] \\ &= \rho(\bar{x}) + \left\{\zeta_x + \sum_{j=1}^{L}(\bar{\lambda}_j^T\mathbf{Q}A_j + \bar{\mu}_j^T\mathbf{F}A_j)\right\}(x-\bar{x}),\end{aligned}$$

where the last equation uses (B.3). The above inequality shows that

$$\xi = \zeta_x + \sum_{j=1}^{L}(\bar{\lambda}_j^T\mathbf{Q}A_j + \bar{\mu}_j^T\mathbf{F}A_j)$$

is a subgradient of ρ at $\bar{x}$.

REFERENCES

[1] H. Benitez-Silva, G. Hall, G. J. Hitsch, G. Pauletto, and J. Rust, *A Comparison of Discrete and Parametric Approximation Methods for Continuous-State Dynamic Programming problems*, Reseach report, Department of Economics, Yale University, New Haven, CT, 2000 (also Working Paper No. 24, Computation in Economics and Finance, Society for Computational Economics, 2000).

[2] D. P. Bertsekas, *Dynamic Programming and Stochastic Control*, Academic Press, New York, 1976.

[3] D. P. Bertsekas, *Dynamic Programming and Optimal Control, Volume* II, Athena Scientific, Belmont, MA, 1995.

[4] D. P. De Farias and B. Van Roy, *The linear programming approach to approximate dynamic programming*, Oper. Res., 51 (2003), pp. 850–865.

[5] R. Horst, P. M. Pardalos, and N. V. Thoai, *Introduction to Global Optimization*, 2nd ed., Kluwer Academic Publishers, Dordrecht, 2000.

[6] R. Horst and H. Tuy, *Global Optimization: Deterministic Approaches*, 2nd revised ed., Springer-Verlag, Heidelberg, 1993.

[7] L. A. Korf, *Approximating infinite horizon stochastic optimal control in discrete time with constraints*, Ann. Oper. Res., 142 (2006), pp. 165–186.

[8] W. S. Lovejoy, *A survey of algorithmic methods for partially observed Markov decision processes*, Ann. Oper. Res., 28 (1991), pp. 47–66.

[9] W. S. Lovejoy, *Computationally feasible bound for partially observed Markov decision processes*, Oper. Res., 39 (1991), pp. 162–175.

[10] R. C. Merton, *Lifetime portfolio selection under uncertainty: The continuous-time case*, Rev. Econom. Stat., 51 (1969), pp. 247–257.

[11] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[12] W. Rudin, *Principles of Mathematical Analysis*, 3rd ed., McGraw-Hill, New York, 1984.

[13] J. Rust, *A Comparison of Policy Iteration Methods for Solving Continuous-State, Infinite-Horizon Markovian Decision Problems using Random, Quasi-Random, and Deterministic Discretizations*, Working Paper, Yale University, April, 1997 (also in Computational Economics, *Economics Working Paper Archive at WUSTL*).

[14] P. A. Samuelson, *Lifetime portfolio selection by dynamic stochastic programming*, Rev. Econom. Stat., 51 (1969), pp. 239–246.

[15] R. D. Smallwood and E. J. Sondik, *The optimal control of partially observable Markov processes over a finite horizon*, Oper. Res., 21 (1973), pp. 1071–1088.

SIAM J. OPTIM.
Vol. 18, No. 4, pp. 1187–1208

THE ADAPTIVE CONVEXIFICATION ALGORITHM: A FEASIBLE POINT METHOD FOR SEMI-INFINITE PROGRAMMING*

CHRISTODOULOS A. FLOUDAS† AND OLIVER STEIN‡

Abstract. We present a new numerical solution method for semi-infinite optimization problems. Its main idea is to adaptively construct convex relaxations of the lower level problem, replace the relaxed lower level problems equivalently by their Karush–Kuhn–Tucker conditions, and solve the resulting mathematical programs with complementarity constraints. This approximation produces *feasible iterates* for the original problem. The convex relaxations are constructed with ideas from the αBB method of global optimization. The necessary upper bounds for second derivatives of functions on box domains can be determined using the techniques of interval arithmetic, where our algorithm already works if only one such bound is available for the problem. We show convergence of stationary points of the approximating problems to a stationary point of the original semi-infinite problem within arbitrarily given tolerances. Numerical examples from Chebyshev approximation and design centering illustrate the performance of the method.

Key words. semi-infinite programming, αBB, global optimization, convex optimization, mathematical program with complementarity constraints, bilevel optimization

AMS subject classifications. 90C34, 90C33, 90C26, 65K05

DOI. 10.1137/060657741

1. Introduction.

We consider the optimization problem

$$\text{SIP}: \qquad \min_{x\in X} f(x) \quad \text{subject to} \quad g(x,y) \le 0 \text{ for all } y \in Y,$$

with objective function $f \in C^2(\mathbb{R}^n, \mathbb{R})$, constraint functions $g \in C^2(\mathbb{R}^n \times \mathbb{R}^m, \mathbb{R}^p)$, a box constraint set $X = [x^\ell, x^u] \subset \mathbb{R}^n$, with $x^\ell < x^u \in \mathbb{R}^n$, and a box index set $Y = [y^\ell, y^u] \subset \mathbb{R}^m$, with $y^\ell < y^u \in \mathbb{R}^m$, where the vector inequalities are understood componentwise. Problems of this type, in which a finite-dimensional decision variable is subject to infinitely many inequality constraints, are called (standard) semi-infinite. In this article we will focus our attention on the simplest case $p = 1$ (one semi-infinite constraint) and $m = 1$; that is, without loss of generality we put $Y = [0, 1]$. More general problem formulations allow Y to be a nonempty compact subset of $\mathbb{R}^m$ or even to be nonfixed but x-dependent. In the latter case the problem is called generalized semi-infinite. Moreover, in applications additional inequality and equality constraints may be present. Whereas the techniques and results presented in this article may be used to tackle these more general problems (with more or less additional effort), for the sake of clarity we develop the main ideas only for SIP with $p = 1$ and $Y = [0, 1]$.

Standard semi-infinite problems have been studied systematically since the 1960s. Important early contributions regarding optimality conditions and duality theory for standard semi-infinite problems are given in [6, 15] for linear semi-infinite problems and in [20, 49] for nonlinear problems. For excellent reviews with hundreds of references on semi-infinite programming, we refer to [21, 36]. A standard reference for

*Received by the editors April 20, 2006; accepted for publication (in revised form) May 10, 2007; published electronically October 10, 2007.
http://www.siam.org/journals/siopt/18-4/65774.html
†Department of Chemical Engineering, Princeton University, Princeton, NJ 08540 (floudas@titan.princeton.edu).
‡Corresponding author. School of Economics and Business Engineering, University of Karlsruhe, 76128 Karlsruhe, Germany (stein@wior.uni-karlsruhe.de). This author gratefully acknowledges support through a Heisenberg grant of the *Deutsche Forschungsgemeinschaft*.

linear semi-infinite problems is [16], and [23, 37, 38] overview the existing numerical methods. Generalized semi-infinite optimization is treated in detail in [45].

Until recently the numerical methods for SIP suffered from the major drawback that their approximations of the feasible set $X \cap M$, with

$$M \;=\; \{\, x \in \mathbb{R}^n |\; g(x,y) \le 0 \text{ for all } y \in Y \,\},$$

may contain infeasible points. In fact, discretization and exchange methods approximate M by finitely many inequalities corresponding to finitely many indices in Y, yielding an outer approximation of M, and reduction-based methods solve the Karush–Kuhn–Tucker system of SIP (cf. section 2.1) by a Newton-SQP approach. As a consequence, the iterates of these methods are not necessarily feasible for SIP, but only their limit might be. Even to ensure the latter, active index sets of the iterates have to be determined. As we will explain, this amounts to the solution of certain global optimization problems.

The recent articles [4, 5] present a branch-and-bound framework for the global solution of SIP which generates convergent sequences of lower and upper bounds for the globally optimal value. Whereas the lower bounds are obtained by usual discretization, the upper bounds are computed using inclusion bounds and interval arithmetic. Whereas [4, 5] focus on the global solution of SIP, they also present the first algorithm with feasible iterates for SIP, to our knowledge.

In fact, checking feasibility of a given point $\bar{x} \in \mathbb{R}^n$ is the crucial problem in semi-infinite optimization. Clearly we have $\bar{x} \in M$ if and only if $\varphi(\bar{x}) \le 0$ holds with the function

$$\varphi :\; \mathbb{R}^n \to \mathbb{R},\; x \mapsto \max_{y \in Y} g(x,y)\,.$$

The latter function is the optimal value function of the so-called lower level problem of SIP,

$$Q(x): \qquad \max_{y \in \mathbb{R}} g(x,y) \quad \text{subject to} \quad 0 \le y \le 1\,.$$

The difficulty lies in the fact that $\varphi(\bar{x})$ is the *globally* optimal value of $Q(\bar{x})$ which might be hard to determine numerically. In fact, standard NLP solvers can only be expected to produce a *local* maximizer y_{loc} of $Q(\bar{x})$ which is not necessarily a global maximizer y_{glob}. Even if $g(\bar{x}, y_{\text{loc}}) \le 0$ is satisfied, $\bar{x}$ might be infeasible since $g(\bar{x}, y_{\text{loc}}) \le 0 < \varphi(\bar{x}) = g(\bar{x}, y_{\text{glob}})$ may hold.

To avoid this effect one could simply assume that the lower level is a convex optimization problem, since then its local and global maximizers coincide. As the feasible set $Y = [0,1]$ of $Q(x)$ is convex, only concavity of g with respect to y for each $x \in X$ is needed to ensure lower level convexity.

The known numerical methods for standard semi-infinite optimization have not taken this approach since the concavity assumption on g fails in most applications. In generalized semi-infinite optimization, however, a number of real-life applications with convex lower level problems exist, so that this structure was exploited in [40, 45, 48]. We will recall the main ideas in section 2.2.

In this article we aim at constructing a sequence of convexifications of the lower level problem and using the techniques from [45, 48] to solve the auxiliary problems with convex lower levels. The convexifications will be produced using the ideas of the αBB method [1, 2, 12] which is explained in section 2.3. As αBB allows one to construct concave overestimators of g, all iterates of our method will be *feasible* for SIP.

Note that in this article we are mainly interested in the feasibility of the iterates, that is, in global solutions of the lower level problem, and *not* in a possibly global solution of SIP (the upper level problem). For this reason we neither make any global assumptions on the defining functions of the upper level such as linearity or convexity with respect to x, nor do we convexify the problem or branch-and-bound with respect to x. Consequently our solution concept will be that of local minimizers or stationary points. This is a main difference to the branch-and-bound approach presented in [4, 5]. While we cannot guarantee global solutions of SIP, our approach can easily be combined with other methods, should a special structure of SIP with respect to x be present. If no special structure is present, our method promises to compute local minimizers with less numerical effort than a method aiming at global minimizers. We point out that our procedure involves the solution of auxiliary optimization problems involving equality constraints, whereas the approach from [4, 5] generally leads to auxiliary problems with nonsmooth inequality constraints.

Section 3 develops our new solution method, Algorithm 3.4, and section 4 provides corresponding convergence results. Numerical examples illustrate the performance of the method in section 5. Finally, in section 6 we point out possible improvements and generalizations of our method as well as some open questions.

2. Background.

2.1. Stationarity conditions. First order necessary optimality conditions for SIP and a natural constraint qualification are well-known and briefly recalled in what follows. To keep the exposition concise, throughout this paper we assume that the box X is so large that the considered stationary points are contained in its interior. Since φ is continuous [9], M is a closed set, and a feasible point $\bar{x}$, with $\varphi(\bar{x}) < 0$, lies in the interior of M. A local minimizer $\bar{x}$ of SIP, with $\varphi(\bar{x}) < 0$, hence necessarily satisfies $\nabla f(\bar{x}) = 0$. Here and in the following, the column vector ∇F denotes the gradient of a C^1 function F and $\nabla_z F$ the gradient of F with respect to the vector z.

For $\bar{x} \in \partial M$, the topological boundary of M, we define the active index set

$$Y_0(\bar{x}) = \{y \in Y | g(\bar{x}, y) = 0\}.$$

Note that $Y_0(\bar{x})$ is nonempty and compact and that it coincides with the set of global maximizers of $Q(\bar{x})$. The following theorem is due to John. For our formulation we use the $(n+1)$-dimensional standard simplex

$$S^{n+1} = \left\{ s \in \mathbb{R}^{n+2} \middle| s \geq 0, \sum_{i=1}^{n+2} s_i = 1 \right\}.$$

THEOREM 2.1 (see [25]). *Let $\bar{x} \in \partial M$ be a local minimizer of* SIP. *Then there exist $y^k \in Y_0(\bar{x})$, $1 \leq k \leq n+1$, and $(\kappa, \lambda) \in \mathbb{R} \times \mathbb{R}^{n+1}$, with $(\kappa, \lambda) \in S^{n+1}$, and*

$$\kappa \nabla f(\bar{x}) + \sum_{k=1}^{n+1} \lambda_k \nabla_x g(\bar{x}, y^k) = 0, \tag{2.1}$$

$$\lambda_k \cdot g(\bar{x}, y^k) = 0, 1 \leq k \leq n+1. \tag{2.2}$$

A point $\bar{x} \in M$ is said to satisfy the *extended Mangasarian–Fromovitz constraint qualification (EMFCQ)* if

$$\nabla_x^\top g(\bar{x}, y) d < 0 \quad \text{for all} \quad y \in Y_0(\bar{x})$$

holds with some vector $d \in \mathbb{R}^n$. Under EMFCQ at $\bar{x}$, one can choose $\kappa = 1$ in (2.1) and thus obtain a *Karush–Kuhn–Tucker* condition [23]. Note that, in the case when strictly less than $n+1$ active indices are sufficient for stationarity, we can satisfy (2.1), (2.2) by artificially listing the same active index an appropriate number of times and setting the corresponding multipliers to zero.

2.2. Convex lower level problems. Assume for the moment that $Q(x)$ is actually a convex optimization problem for all $x \in X$; that is, $g(x, \cdot)$ is concave on $Y = [0, 1]$ for these x. The approach from [45, 48] then takes advantage of the fact that the solution set of a differentiable convex lower level problem satisfying a constraint qualification is characterized by its first order optimality condition. As a first step for this, observe that SIP and the Stackelberg game

$$\text{SG}: \quad \min_{x,y} f(x) \quad \text{subject to} \quad g(x, y) \leq 0, \text{ and } y \text{ solves } Q(x),$$

are equivalent problems (this can even be shown for generalized semi-infinite problems; cf. [47]). Next, the restriction "y solves $Q(x)$" in SG can be equivalently replaced by its Karush–Kuhn–Tucker condition. For this reformulation we use that the Lagrange function of $Q(x)$

$$\mathcal{L}(x, y, \gamma_\ell, \gamma_u) = g(x, y) + \gamma_\ell y + \gamma_u(1 - y)$$

satisfies

$$\nabla_y \mathcal{L}(x, y, \gamma_\ell, \gamma_u) = \nabla_y g(x, y) + \gamma_\ell - \gamma_u$$

and obtain that the Stackelberg game is equivalent to the following mathematical program with complementarity constraints:

$$\begin{aligned} \text{MPCC}: \quad \min_{x,y,\gamma_\ell,\gamma_u} f(x) \quad \text{subject to} \quad & g(x, y) \leq 0 \\ & \nabla_y g(x, y) + \gamma_\ell - \gamma_u = 0 \\ & \gamma_\ell y = 0 \\ & \gamma_u(1 - y) = 0 \\ & \gamma_\ell, \gamma_u \geq 0 \\ & y, 1 - y \geq 0. \end{aligned}$$

Unfortunately standard numerical software cannot be expected to solve this problem since due to the appearance of complementarity conditions the Mangasarian-Fromovitz constraint qualification (MFCQ) is violated at all points of the feasible set of MPCC [42]. In [27, 39] it is shown that MFCQ is a necessary condition for the stability of smooth nonlinear programs under data perturbations and thus for the stability of numerical methods in the presence of roundoff errors. Note that in [11] it is shown, however, that at least certain SQP methods will converge locally for MPCC. For overviews of MPCC solution methods we refer to [29, 30, 43].

One approach to solve MPCC is the reformulation of the complementarity constraints by a so-called NCP function, that is, a function $\phi : \mathbb{R}^2 \to \mathbb{R}$, with

$$\phi(a, b) = 0 \quad \text{if and only if} \quad a \geq 0, b \geq 0, ab = 0.$$

Examples are the natural residual or min-function

$$\phi^{NR}(a, b) = \frac{1}{2}\left(a + b - \sqrt{(a - b)^2}\right),$$

and the Fischer–Burmeister function [10]

$$\phi^{FB}(a,b) = a + b - \sqrt{a^2 + b^2}.$$

For numerical purposes one can regularize these nondifferentiable NCP functions. The so-called Chen–Harker–Kanzow–Smale function [7, 28, 44] is given by

$$\phi_\tau^{NR}(a,b) = \frac{1}{2}\left(a + b - \sqrt{(a-b)^2 + 4\tau^2}\right),$$

whereas the so-called smoothed Fischer–Burmeister function is

$$\phi_\tau^{FB}(a,b) = a + b - \sqrt{a^2 + b^2 + 2\tau^2}.$$

Obviously ϕ_τ^{NR} and ϕ_τ^{FB} are continuously differentiable for all $\tau \neq 0$, and for $\tau = 0$ they coincide with ϕ^{NR} and ϕ^{FB}, respectively.

With ϕ denoting one of the NCP functions ϕ^{NR} and ϕ^{FB}, MPCC can now be equivalently rewritten as the nonsmooth problem

$$\begin{aligned} P: \quad \min_{x,y,\gamma_\ell,\gamma_u} \ f(x) \quad \text{subject to} \quad & g(x,y) \le 0 \\ & \nabla_y g(x,y) + \gamma_\ell - \gamma_u = 0 \\ & \phi(\gamma_\ell, y) = 0 \\ & \phi(\gamma_u, 1-y) = 0, \end{aligned}$$

where, in particular, the inequality constraints on y, γ_ℓ, γ_u are now modeled by ϕ. Problem P is embedded into the parameterized family of optimization problems

$$\begin{aligned} P_\tau: \quad \min_{x,y,\gamma_\ell,\gamma_u} \ f(x) \quad \text{subject to} \quad & g(x,y) \le 0 \\ & \nabla_y g(x,y) + \gamma_\ell - \gamma_u = 0 \\ & \phi_\tau(\gamma_\ell, y) = 0 \\ & \phi_\tau(\gamma_u, 1-y) = 0, \end{aligned}$$

which are smooth for $\tau \neq 0$ and satisfy $P_0 = P$. Given a sequence of regularization parameters $\tau^\nu \searrow 0$, the problems P_{τ^ν} can be solved by standard software, and in [45, 48] it is shown that under weak assumptions each cluster point $x^\star$ of the solution components x^ν is a solution of SIP.

2.3. The αBB method. In αBB, a convex underestimator of a nonconvex function is constructed by decomposing it into a sum of nonconvex terms of special type (e.g., linear, bilinear, trilinear, fractional, fractional trilinear, convex, univariate concave) and nonconvex terms of arbitrary type. The first type is then replaced by its convex envelope or very tight convex underestimators which are already known. A complete list of the tight convex underestimators of the above special-type nonconvex terms is provided in [12]. Recent comprehensive reviews on the state of the art in global optimization can be found [13, 14].

To keep the exposition concise, in this article we will treat all terms as arbitrarily nonconvex. For these terms, αBB constructs convex underestimators by adding a quadratic relaxation function ψ. With the obvious modification we use this approach to construct a concave overestimator for a nonconcave function $g : [y^\ell, y^u] \to \mathbb{R}$ being C^2 on an open neighborhood of $[y^\ell, y^u]$. With

$$\psi(y; \alpha, y^\ell, y^u) = \tfrac{\alpha}{2}(y - y^\ell)(y^u - y), \tag{2.3}$$

we put

$$\tilde{g}(y;\alpha,y^\ell,y^u) = g(y) + \psi(y;\alpha,y^\ell,y^u).$$

In what follows we will suppress the dependence of $\tilde{g}$ on y^ℓ, y^u. For $\alpha \geq 0$ the function $\tilde{g}$ clearly is an overestimator of g on $[y^\ell, y^u]$, and it coincides with g at the end points y^ℓ, y^u of the domain. Moreover, $\tilde{g}$ is twice continuously differentiable with the second derivative

$$\nabla_y^2 \tilde{g}(y;\alpha) = \nabla^2 g(y) - \alpha$$

on $[y^\ell, y^u]$. Consequently $\tilde{g}$ is concave on $[y^\ell, y^u]$ for

$$\alpha \geq \max_{y\in[y^\ell,y^u]} \nabla^2 g(y) \tag{2.4}$$

(cf. also [1, 2]). The computation of α thus involves a global optimization problem itself. Note, however, that we may use *any* upper bound for the right-hand side in (2.4). Such upper bounds can be provided by interval methods (see, e.g., [12, 19, 33]). We will call α satisfying (2.4) a *convexification parameter*.

Combining these facts shows that for

$$\alpha \geq \max\left(0, \max_{y\in[y^\ell,y^u]} \nabla^2 g(y)\right) \tag{2.5}$$

the function $\tilde{g}(y;\alpha)$ is a concave overestimator of g on $[y^\ell, y^u]$. A representative measure of the quality of an overestimator is the separation distance between itself and the nonconcave function it overestimates. The smaller the separation distance, the tighter the overestimator is. The separation distance between $g(y)$ and $\tilde{g}(y;\alpha)$ is defined by the difference of these functions; that is,

$$d_{\alpha BB}(y;\alpha) = \tilde{g}(y;\alpha) - g(y) = \tfrac{\alpha}{2}(y - y^\ell)(y^u - y).$$

The separation distance is hence maximal at the midpoint of the interval $[y^\ell, y^u]$, and its value is

$$\max_{y\in[y^\ell,y^u]} d_{\alpha BB}(y;\alpha) = \tfrac{\alpha}{8}(y^u - y^\ell)^2 \tag{2.6}$$

(see also [32]). On the one hand, (2.6) quantifies the improvement of the maximal separation distance for better bounds α with (2.5) on a fixed set $[y^\ell, y^u]$; on the other hand, it shows that the maximal separation distance decreases quadratically for $|y^u - y^\ell| \to 0$, even if α is bounded below by some positive number.

Whereas the presented ideas are used in the αBB approach to global optimization in a branch-and-bound framework [12], here we will use them for adaptive convexifications of the lower level problem of SIP.

3. The numerical approach. For $N \in \mathbb{N}$ let $0 = \eta^0 < \eta^1 < \cdots < \eta^{N-1} < \eta^N = 1$ define a subdivision of $Y = [0,1]$; that is, with $K = \{1,\ldots,N\}$ and

$$Y^k = [\eta^{k-1}, \eta^k], \quad k \in K,$$

we have

$$Y = \bigcup_{k\in K} Y_k. \tag{3.1}$$

A trivial but very useful observation is that the single semi-infinite constraint

$$g(x,y) \le 0 \quad \text{for all } y \in Y$$

is equivalent to the finitely many semi-infinite constraints

$$g(x,y) \le 0 \quad \text{for all } y \in Y^k, k \in K.$$

Given a subdivision, we will construct concave overestimators for each of these finitely many semi-infinite constraints, solve the corresponding optimization problem, and adaptively refine the subdivision.

The following lemma formulates the obvious fact that replacing g by overestimators on each subdivision node Y^k results in an approximation of M by *feasible* points.

LEMMA 3.1. *For each $k \in K$ let $g^k : X \times Y^k \to \mathbb{R}$, and let $\bar{x} \in X$ be given such that for all $k \in K$ and all $y \in Y^k$ we have $g(\bar{x},y) \le g^k(\bar{x},y)$. Then the constraints*

$$g^k(\bar{x},y) \le 0 \quad \textit{for all } y \in Y^k, k \in K,$$

entail $\bar{x} \in M$.

Proof. The proof immediately follows from (3.1). ∎

3.1. αBB for the lower level. For the construction of these overestimators we use ideas of the αBB method. In fact, for each $k \in K$ we put

$$g^k : X \times Y^k \to \mathbb{R}, (x,y) \mapsto g(x,y) + \psi(y; \alpha_k, \eta^{k-1}, \eta^k) \tag{3.2}$$

with the quadratic relaxation function ψ from (2.3) and

$$\alpha_k > \max\left(0, \max_{(x,y) \in X \times Y^k} \nabla_y^2 g(x,y)\right). \tag{3.3}$$

Note that the latter condition on α_k is uniform in x. We emphasize that with the single bound

$$\bar{\alpha} > \max\left(0, \max_{(x,y) \in X \times Y} \nabla_y^2 g(x,y)\right) \tag{3.4}$$

the choices $\alpha_k := \bar{\alpha}$ satisfy (3.3) for all $k \in K$. Moreover, the α_k can always be chosen such that $\alpha_k \le \bar{\alpha}$, $k \in K$.

The following properties of g^k are easily verified.

LEMMA 3.2. *For each $k \in K$ let g^k be given by (3.2). Then the following holds:*

(i) *For all $(x,y) \in X \times Y^k$ we have $g(x,y) \le g^k(x,y)$.*
(ii) *For all $x \in X$, the function $g^k(x,\cdot)$ is concave on Y^k.*
(iii) *The maximal separation distance (compare (2.6)) between g^k and g on $X \times Y^k$ is $\frac{\alpha_k}{8}(\eta^k - \eta^{k-1})^2$.*

Now we consider the following approximation of the feasible set M, where $E = \{\eta^k | k \in K\}$ denotes the set of subdivision points and α the vector of convexification parameters:

$$M_{\alpha BB}(E,\alpha) = \{x \in \mathbb{R}^n | g^k(x,y) \le 0 \quad \text{for all } y \in Y^k, k \in K\}.$$

By Lemmas 3.1 and 3.2(i) we have $M_{\alpha BB}(E,\alpha) \subset M$. This means that any solution concept for

$$\mathrm{SIP}_{\alpha BB}(E,\alpha): \quad \min_{x \in X} f(x) \quad \text{subject to} \quad x \in M_{\alpha BB}(E,\alpha),$$

be it global solutions, local solutions or stationary points, will at least lead to feasible points of SIP (provided that $\mathrm{SIP}_{\alpha BB}(E,\alpha)$ is consistent, i.e., its feasible set is nonvoid).

The problem $\mathrm{SIP}_{\alpha BB}(E,\alpha)$ has finitely many lower level problems $Q^k(x)$, $k \in K$, with

$$Q^k(x): \qquad \max_{y\in\mathbb{R}} g^k(x,y) \quad \text{subject to} \quad \eta^{k-1} \le y \le \eta^k.$$

Since the inequality (3.3) is strict, the convex problem $Q^k(x)$ has a unique solution $y^k(x)$ for each $k \in K$ and $x \in X$. Note that the functions y^k, $k \in K$, are even Lipschitz continuous on X, since the strong second order sufficiency condition and the linear independence constraint qualification hold at $y^k(x)$ [39].

Recall that $y \in Y^k$ is called active for the constraint $\max_{y\in Y^k} g^k(x,y) \le 0$ at $\bar{x}$ if $g^k(\bar{x},y) = 0$ holds. By the uniqueness of the global solution of $Q^k(\bar{x})$ there exists *at most one* active index for each $k \in K$, namely, $y^k(\bar{x})$. Thus, we can consider the finite active index sets

$$\begin{aligned} K_0(\bar{x}) &= \{k \in K | g^k(\bar{x}, y^k(\bar{x})) = 0\},\\ Y_0^{\alpha BB}(\bar{x}) &= \{y^k(\bar{x}) | k \in K_0(\bar{x})\}. \end{aligned}$$

3.2. The MPCC reformulation. Following the ideas of section 2.2 we find that y^k solves $Q^k(x)$ if and only if $(x, y^k, \gamma_\ell^k, \gamma_u^k)$ solves the system

$$\begin{aligned} \nabla_y g^k(x,y) + \gamma_\ell - \gamma_u &= 0,\\ \phi(\gamma_\ell, y - \eta^{k-1}) &= 0,\\ \phi(\gamma_u, \eta^k - y) &= 0, \end{aligned}$$

with some γ_ℓ^k, γ_u^k, and ϕ denoting one of the NCP functions ϕ^{NR} and ϕ^{FB}. With

$$\begin{aligned} w &:= (x, y^k, \gamma_\ell^k, \gamma_u^k, k \in K),\\ F(w) &:= f(x),\\ G^k(w;E,\alpha) &:= g(x,y^k) + \tfrac{\alpha_k}{2}(y^k - \eta^{k-1})(\eta^k - y^k),\\ H^k(w;E,\alpha) &:= \begin{pmatrix} \nabla_y g(x,y^k) + \alpha_k\left(\frac{\eta^{k-1}+\eta^k}{2} - y^k\right) + \gamma_\ell^k - \gamma_u^k \\ \phi(\gamma_\ell^k, y^k - \eta^{k-1}) \\ \phi(\gamma_u^k, \eta^k - y^k) \end{pmatrix}, \end{aligned}$$

we can thus replace $\mathrm{SIP}_{\alpha BB}(E,\alpha)$ equivalently by the nonsmooth problem

$$P(E,\alpha): \qquad \min_w F(w) \qquad \text{subject to} \qquad \begin{aligned} &G^k(w;E,\alpha) \le 0,\\ &H^k(w;E,\alpha) = 0, k \in K. \end{aligned}$$

The latter problem can be solved numerically by regularization of the NCP functions; that is, H^k is replaced by

$$H^k(w;E,\alpha,\tau) = \begin{pmatrix} \nabla_y g(x,y^k) + \alpha_k\left(\frac{\eta^{k-1}+\eta^k}{2} - y^k\right) + \gamma_\ell^k - \gamma_u^k \\ \phi_\tau(\gamma_\ell^k, y^k - \eta^{k-1}) \\ \phi_\tau(\gamma_u^k, \eta^k - y^k) \end{pmatrix},$$

and for $\tau \neq 0$ we obtain the smooth problem

$$P(E,\alpha,\tau): \qquad \min_w F(w) \qquad \text{subject to} \qquad \begin{aligned} &G^k(w;E,\alpha) \le 0,\\ &H^k(w;E,\alpha,\tau) = 0, k \in K. \end{aligned}$$

It may be tackled by numerical standard software which we will consider to be a "black box." One possibility is to compute solutions of these problems for a user-defined sequence $\tau^\nu \searrow 0$, as suggested in [45, 48]. In view of the simple structure of the complementarity constraints, in the present setting one can also try to let the NLP solver itself drive τ from a positive initial value to zero while solving the problem

$$\tilde{P}(E,\alpha): \quad \min_{(w,\tau)} F(w) \quad \text{subject to} \quad \begin{aligned} &G^k(w;E,\alpha) \le 0,\\ &H^k(w;E,\alpha,\tau) = 0, k \in K,\\ &\tau = 0. \end{aligned}$$

For our numerical examples in section 5 this alternative works well. In any case we assume that the black box NLP solver generates a local minimizer $\bar{w}$ of $P(E,\alpha)$. The subvector $\bar{x}$ of $\bar{w}$ is then a local minimizer and, hence, a stationary point of $\mathrm{SIP}_{\alpha BB}(E,\alpha)$. Note that in the present article we do not discuss the effects of numerical inaccuracies in the solution of $P(E,\alpha)$ on the feasibility of iterates for SIP.

3.3. The adaptive convexification algorithm. The main idea of our numerical method is to compute a stationary point $\bar{x}$ of $\mathrm{SIP}_{\alpha BB}(E,\alpha)$ by the approach from section 3.2 and terminate if $\bar{x}$ is also stationary for SIP within given tolerances. If $\bar{x}$ is not stationary we refine the subdivision E in the spirit of exchange methods [21, 38] by adding the active indices $Y_0^{\alpha BB}(\bar{x})$ to E and construct a refined problem $\mathrm{SIP}_{\alpha BB}(E \cup Y_0^{\alpha BB}(\bar{x}), \tilde{\alpha})$ by the following procedure. Note that, in view of Carathéodory's theorem (compare also Theorem 2.1), the number of elements of $Y_0^{\alpha BB}(\bar{x})$ may be bounded by $n+1$.

Refinement step:

For any $\tilde{\eta} \in Y_0^{\alpha BB}(\bar{x})$, let $k \in K$ be the index with $\tilde{\eta} \in [\eta^{k-1}, \eta^k]$. Put $Y^{k,1} = [\eta^{k-1}, \tilde{\eta}]$, $Y^{k,2} = [\tilde{\eta}, \eta^k]$, let $\alpha_{k,1}$ and $\alpha_{k,2}$ be the corresponding convexification parameters, put

$$\begin{aligned} g^{k,1}(x,y) &= g(x,y) + \frac{\alpha_{k,1}}{2}(y-\eta^{k-1})(\tilde{\eta}-y),\\ g^{k,2}(x,y) &= g(x,y) + \frac{\alpha_{k,2}}{2}(y-\tilde{\eta})(\eta^k-y), \end{aligned}$$

and define $M_{\alpha BB}(E \cup \{\tilde{\eta}\}, \tilde{\alpha})$ by replacing the constraint

$$g^k(x,y) \le 0 \quad \text{for all } y \in Y^k$$

in $M_{\alpha BB}(E,\alpha)$ by the two new constraints

$$g^{k,i}(x,y) \le 0 \quad \text{for all } y \in Y^{k,i}, i = 1,2,$$

and by replacing the entry α_k of α by the two new entries $\alpha_{k,i}$, $i = 1,2$.

The point $\bar{x}$ is stationary for $\mathrm{SIP}_{\alpha BB}(E,\alpha)$ (in the sense of John; compare Theorem 2.1) if $\bar{x} \in M_{\alpha BB}(E,\alpha)$ and if there exist $y^k \in Y_0^{\alpha BB}(\bar{x})$, $1 \le k \le n+1$, and $(\kappa,\lambda) \in S^{n+1}$ with

$$\kappa \nabla f(\bar{x}) + \sum_{k=1}^{n+1} \lambda_k \nabla_x g(\bar{x}, y^k) = 0, \tag{3.5}$$

$$\lambda_k \cdot g^k(\bar{x}, y^k) = 0, 1 \le k \le n+1. \tag{3.6}$$

Note that in (3.5) we simplified

$$\nabla_x g^k(\bar{x}, y^k) = \nabla_x \left(g(x, y^k) + \psi(y^k; \alpha_k, \eta^{k-1}, \eta^k) \right) = \nabla_x g(x, y^k).$$

For the numerical algorithm it is crucial to relax the notions of *active index*, *stationarity*, and *set unification* by certain tolerances.

DEFINITION 3.3. *For $\varepsilon_{act}, \varepsilon_{stat}, \varepsilon_{\cup} > 0$ we say that*

(i) *y^k is ε_{act}-*active for g^k at $\bar{x}$ *if $g^k(\bar{x}, y^k) \in [-\varepsilon_{act}, 0]$,*

(ii) *$\bar{x}$ is ε_{stat}-*stationary for SIP with *ε_{act}-*active indices *if $\bar{x} \in M$ and if there exist $y^k \in Y$, $1 \le k \le n+1$, and $(\kappa, \lambda) \in S^{n+1}$ such that*

$$\left\| \kappa \nabla f(\bar{x}) + \sum_{k=1}^{n+1} \lambda_k \nabla_x g(\bar{x}, y^k) \right\| \le \varepsilon_{stat}, \tag{3.7}$$

$$\lambda_k \cdot g(\bar{x}, y^k) \in [-\lambda_k \cdot \varepsilon_{act}, 0],\ 1 \le k \le n+1, \tag{3.8}$$

hold, and

(iii) *the $\varepsilon_{\cup}$-*union of E and $\tilde{\eta}$ *is $E \cup \{\tilde{\eta}\}$ if*

$$\min\{\tilde{\eta} - \eta^{k-1}, \eta^k - \tilde{\eta}\} > \varepsilon_{\cup} \cdot (\eta^k - \eta^{k-1})$$

holds for the $k \in K$ with $\tilde{\eta} \in [\eta^{k-1}, \eta^k]$, and E otherwise (i.e., $\tilde{\eta}$ is not unified with E if its distance from E is too small).

ALGORITHM 3.4 (adaptive convexification algorithm).

Step 1: *Determine a uniform convexification parameter $\bar{\alpha}$ with* (3.4)*, choose $N \in \mathbb{N}$, $\eta^k \in Y$, and $\alpha_k \le \bar{\alpha}$, $k \in K = \{1, \dots, N\}$, such that* $\mathrm{SIP}_{\alpha BB}(E, \alpha)$ *is consistent, as well as tolerances $\varepsilon_{act}, \varepsilon_{stat}, \varepsilon_{\cup} > 0$, with $\varepsilon_{\cup} \le 2\varepsilon_{act}/\bar{\alpha}$.*

Step 2: *By solving $P(E, \alpha)$, compute a stationary point x of* $\mathrm{SIP}_{\alpha BB}(E, \alpha)$ *with ε_{act}-active indices y^k, $1 \le k \le n+1$, and multipliers (κ, λ).*

Step 3: *Terminate if x is ε_{stat}-stationary for* SIP *with $(2\varepsilon_{act})$-active indices y^k, $1 \le k \le n+1$, from Step 2 and multipliers (κ, λ) from Step 2.*
Otherwise construct a new set $\tilde{E}$ of subdivision points as the $\varepsilon_{\cup}$-union of E and $\{y^k | 1 \le k \le n+1\}$ and perform a refinement step for the elements in $\tilde{E} \setminus E$ to construct a new feasible set $M_{\alpha BB}(\tilde{E}, \tilde{\alpha})$.

Step 4: *Put $E = \tilde{E}$, $\alpha = \tilde{\alpha}$, and go to Step 2.*

In section 4 we will show that Algorithm 3.4 is well-defined, convergent, and finitely terminating. Note that after its termination we can exploit that the set $E \subset Y$ contains indices that should also yield a good *outer* approximation of M. The optimal value of the problem

$$P_{outer}: \qquad \min_{x \in X} f(x) \quad \text{subject to} \quad g(x, \eta) \le 0, \eta \in E,$$

yields a rigorous *lower* bound for the optimal value of SIP. If P_{outer} can actually be solved to global optimality (e.g., if a standard NLP solver is used, due to convexity with respect to x), then a comparison of this lower bound for the optimal value of SIP with the upper bound from Algorithm 3.4 can yield a certificate of global optimality for SIP (see also section 5).

3.4. A consistent initial approximation. Even if the feasible set M of SIP is consistent, there is of course no guarantee that its approximations $M_{\alpha BB}(E, \alpha)$ are also consistent. For Step 1 of Algorithm 3.4 we thus suggest the following phase I approach: Use Algorithm 3.4 to construct adaptive convexifications of

$$\mathrm{SIP}^{ph.I}: \qquad \min_{(x,z) \in X \times \mathbb{R}} z \quad \text{subject to} \quad g(x, y) \le z \text{ for all } y \in Y$$

until a feasible point $(\bar{x}, \bar{z})$, with $\bar{z} \le 0$, of $\mathrm{SIP}^{ph.I}_{\alpha BB}(E, \alpha)$ is found with some subdivision E and convexification parameters α. The point $\bar{x}$ is then obviously also feasible for $\mathrm{SIP}_{\alpha BB}(E, \alpha)$ and can be used as an initial point to solve the latter problem. Due to the possible nonconvexity of the upper level problem of SIP, this phase I approach is not necessarily successful, but in our numerical examples in section 5 it works well. If it does not work although SIP is consistent, one can try to initialize phase I with different starting values or use an exhaustive subdivision of the index set like in the subsequent Proposition 4.1.

To initialize Algorithm 3.4 for phase I, select some point $\bar{x}$ in the box X and put $E^1 = \{0, 1\}$, that is, $Y^1 = Y = [0, 1]$. Compute α_1 according to (3.3), and solve the convex optimization problem $Q^1(\bar{x})$ with standard software. With its optimal value $\bar{z}$, the point $(\bar{x}, \bar{z})$ is feasible for $\mathrm{SIP}^{ph.I}_{\alpha BB}(E^1, \alpha_1)$.

4. Convergence results. To obtain a first impression of the approximation properties of our approach, for the following result we temporarily assume that the subdivision of $Y = [0, 1]$ is not adaptive in each step, but equidistant, that is, $\eta^k = k/N$, $k \in K = \{0, \dots, N\}$, with some $N \in \mathbb{N}$. Owing to this exhaustion property, each interior point $\bar{x}$ of M with $\varphi(\bar{x}) < 0$ is feasible for $\mathrm{SIP}_{\alpha BB}(E, \alpha)$ for sufficiently large N (see also [5]). Moreover, boundary points of M are "approximated at a quadratic rate."

PROPOSITION 4.1. *Let $\eta^k = k/N$, and choose $\alpha_k \le \bar{\alpha}$ according to* (3.4)*, $k \in K$.*

(i) *For $\bar{x} \in M$, with $\varphi(\bar{x}) < 0$, let*

$$N \ge \frac{1}{2}\sqrt{\frac{\max_{k \in K} \alpha_k}{|\varphi(\bar{x})|}}. \tag{4.1}$$

Then $\bar{x}$ is feasible for $\mathrm{SIP}_{\alpha BB}(E, \alpha)$.

(ii) *For $\bar{x} \in M$, with $\varphi(\bar{x}) = 0$, the infeasibility measure*

$$\max(0, \max_{k \in K} \max_{y \in Y^k} g^k(\bar{x}, y))$$

of $\bar{x}$ with respect to $M_{\alpha BB}(E, \alpha)$ is of order $O(1/N^2)$.

Proof. For all $k \in K$ and all $y \in Y^k = [k-1, k]/N$, Lemma 3.2(iii) yields

$$\begin{aligned} g^k(\bar{x}, y) &= g(\bar{x}, y) + \frac{\alpha_k}{2}\left(y - \frac{k-1}{N}\right)\left(\frac{k}{N} - y\right) \\ &\le \varphi(\bar{x}) + \frac{\alpha_k}{8N^2} \le \varphi(\bar{x}) + \frac{\max_{k \in K} \alpha_k}{8N^2}. \end{aligned} \tag{4.2}$$

In the case $\varphi(\bar{x}) < 0$ the last term is nonpositive for N with (4.1), which shows the assertion of part (i).

To see part (ii) note that, for $\varphi(\bar{x}) = 0$, (4.2) implies

$$\max_{k \in K} \max_{y \in Y^k} g^k(\bar{x}, y) \le \frac{\bar{\alpha}}{8N^2}$$

and that of course we also have $0 \le \bar{\alpha}/(8N^2)$. ∎

However, in Algorithm 3.4 the subdivision of Y is *not* exhaustive, and we cannot expect the method to approximate the whole interior of M. In particular, the method might not find a global minimizer of SIP contained in a region of M which is not approximated by the method. Recall, on the other hand, that our algorithm is not designed to find global minimizers. Instead of exponential growth of the number of

subdivision points in the exhaustive case, in view of Carathéodory's theorem each of our refinement steps adds at most $n+1$ new points to the subdivision. That is, the growth in the number of new constraints and, thus, in the number of auxiliary variables and auxiliary constraints (cf. section 3.2) is *linear* in n.

Note that Proposition 4.1 will not be used in what follows, where we will focus on adaptive instead of exhaustive subdivisions for the index set. The next results show that even adaptive subdivisions entail at least certain monotonicity properties for the feasible sets and optimal values of the approximating problems.

LEMMA 4.2. *With given E and α, let $\bar{x} \in M_{\alpha BB}(E, \alpha)$. For any additional subdivision point $\tilde{\eta} \notin E$, let $M_{\alpha BB}(E \cup \{\tilde{\eta}\}, \tilde{\alpha})$ be constructed by the refinement step from section* 3.3. *Then we also have $\bar{x} \in M_{\alpha BB}(E \cup \{\tilde{\eta}\}, \tilde{\alpha})$.*

Proof. Let $\tilde{\eta} \in [\eta^{k-1}, \eta^k]$ for some $k \in K$. For $i \in \{1, 2\}$ simple monotonicity arguments show that all $y \in Y^{k,i}$ satisfy $g^{k,i}(\bar{x}, y) \leq g^k(\bar{x}, y) \leq 0$, where the second inequality is due to $Y^{k,i} \subset Y^k$. ∎

PROPOSITION 4.3. *Let $(E^\nu)_\nu$ be a sequence of subdivisions of $Y = [0, 1]$ such that for all $\nu \in \mathbb{N}$ we have $E^\nu \subset E^{\nu+1}$, let the sets $M^\nu := M_{\alpha BB}(E^\nu, \alpha^\nu)$ be constructed by the refinement steps from section* 3.3, *and let v^ν denote the optimal values of $\mathrm{SIP}_{\alpha BB}(E^\nu, \alpha^\nu)$, $\nu \in \mathbb{N}$ (with $v^\nu := +\infty$ for $X \cap M^\nu = \emptyset$). Then the following holds:*

(i) *The feasible sets M^ν satisfy*

$$M^1 \subset M^2 \subset \cdots \subset M^\nu \subset M^{\nu+1} \subset \cdots \subset M.$$

(ii) *If* SIP *is solvable, then the optimal values v^ν converge to an upper bound for the optimal value of* SIP.

(iii) *Each sequence $(x^\nu)_\nu$, with $x^\nu \in X \cap M^\nu$, $\nu \in \mathbb{N}$, has an accumulation point $x^\star$, and all such accumulation points are elements of $X \cap M$.*

Proof. Assertion (i) follows immediately from Lemmas 3.1 and 4.2.

Assertion (ii) is trivial if all sets $X \cap M^\nu$, $\nu \in \mathbb{N}$, are empty. Otherwise, for $X \cap M^{\nu_0} \neq \emptyset$ the sequence $(v^\nu)_{\nu \geq \nu_0}$ is real-valued, monotonically decreasing by part (i), and bounded below by the optimal value of SIP. This shows part (ii).

To see assertion (iii) recall that $X \cap M$ is compact. This implies the existence of an accumulation point $x^\star$. By part (i) we have $x^\nu \in X \cap M$ for all $\nu \in \mathbb{N}$, so that the closedness of $X \cap M$ yields the assertion. ∎

The following corollary is immediate.

COROLLARY 4.4. *Let $(x^\nu)_\nu$ be a sequence of points generated by Algorithm* 3.4. *Then all x^ν, $\nu \in \mathbb{N}$, are feasible for* SIP, *the sequence $(x^\nu)_\nu$ has an accumulation point, each such accumulation point $x^\star$ is feasible for* SIP, *and $f(x^\star)$ provides an upper bound for the optimal value of* SIP.

Next we make sure that Algorithm 3.4 is well defined. The following lemma shows that in Step 3 at least one subdivision point is added to E if the termination criterion is violated.

LEMMA 4.5. *Let Algorithm* 3.4 *be initialized by Step* 1, *and for given E and α in Step* 2 *let x be a stationary point of $\mathrm{SIP}_{\alpha BB}(E, \alpha)$ with ε_{act}-active indices y^k, $1 \leq k \leq n+1$, and multipliers (κ, λ). Then, if the termination criterion in Step* 3 *is violated, the set of new subdivision points $\tilde{E}$ in Step* 3 *is strictly larger than E.*

Proof. Assume $\tilde{E} = E$. By Proposition 4.3(i), x is feasible for SIP, and the stationarity condition (3.5) with certain multipliers implies (3.7) with the same multipliers for arbitrary $\varepsilon_{stat} > 0$.

Since the $y^k \in Y$ are ε_{act}-active, we have

$$g^k(x, y^k) = g(x, y^k) + \frac{\alpha_k}{2}(y^k - \eta^{k-1})(\eta^k - y^k) \in [-\varepsilon_{act}, 0]$$

and thus

$$0 \geq -\frac{\alpha_k}{2}(y^k - \eta^{k-1})(\eta^k - y^k) \geq g(x, y^k) \geq -\varepsilon_{act} - \frac{\alpha_k}{2}(y^k - \eta^{k-1})(\eta^k - y^k).$$

Because of $\tilde{E} = E$, for each k at least one of the terms $y^k - \eta^{k-1}$, $\eta^k - y^k$ is bounded above by $\varepsilon_\cup$, and the other one trivially by 1 (the length of Y). According to Step 1 of the algorithm we also have $\alpha_k \leq \bar{\alpha}$ and $\varepsilon_\cup \leq 2\varepsilon_{act}/\bar{\alpha}$. Combining these estimates yields $g(x, y^k) \in [-2\varepsilon_{act}, 0]$; that is, each y^k, $1 \leq k \leq n+1$, is a $(2\varepsilon_{act})$-active index for x in the original problem SIP.

We have thus shown that x is ε_{stat}-stationary for SIP with $(2\varepsilon_{act})$-active indices y^k, $1 \leq k \leq n+1$, from Step 2 and multipliers (κ, λ) from Step 2. This contradicts the assumption that the termination criterion in Step 3 is violated. ∎

THEOREM 4.6. *Algorithm* 3.4 *terminates after finitely many iterations.*

Proof. Assume that the algorithm does not terminate. Then there exist sequences $(E^\nu)_\nu$, $(\alpha^\nu)_\nu$, $(x^\nu)_\nu$, $(y^{k,\nu})_\nu$, $1 \leq k \leq n+1$, and $(\kappa^\nu, \lambda^\nu)_\nu$ such that for each $\nu \in \mathbb{N}$ the point x^ν is stationary for $\mathrm{SIP}_{\alpha BB}(E^\nu, \alpha^\nu)$ with ε_{act}-active indices $y^{k,\nu}$, $1 \leq k \leq n+1$, and multipliers $(\kappa^\nu, \lambda^\nu)$, but the termination criterion is not satisfied for any $\nu \in \mathbb{N}$.

As the sequence $(x^\nu, y^{1,\nu}, \ldots, y^{n+1,\nu}, \kappa^\nu, \lambda^\nu)_\nu$ is contained in the compact set $X \times Y^{n+1} \times S^{n+1}$, it possesses an accumulation point $(x^\star, y^{1,\star}, \ldots, y^{n+1,\star}, \kappa^\star, \lambda^\star)$ in the same set. By continuity (3.5) yields

$$\kappa^\star \nabla f(x^\star) + \sum_{k=1}^{n+1} \lambda_k^\star \nabla_x g(x^\star, y^{k,\star}) = 0,$$

so that for some $\nu_0 \in \mathbb{N}$ and infinitely many $\nu \geq \nu_0$ we have

$$\left|\left| \kappa^\nu \nabla f(x^\nu) + \sum_{k=1}^{n+1} \lambda_k^\nu \nabla_x g(x^\nu, y^{k,\nu}) \right|\right| < \varepsilon_{stat}.$$

In view of Proposition 4.3 each such point x^ν lies in M, and by the construction in Step 2 we have $g^k(x^\nu, y^{k,\nu}) \in [-\varepsilon_{act}, 0]$, which implies

$$\lambda_k^\nu \cdot g^k(x^\nu, y^{k,\nu}) \in [-\lambda_k^\nu \cdot \varepsilon_{act}, 0], 1 \leq k \leq n+1. \tag{4.3}$$

It remains to be shown that for some $\nu \in \mathbb{N}$ we can replace (4.3) by

$$\lambda_k^\nu \cdot g(x^\nu, y^{k,\nu}) \in [-2\lambda_k^\nu \cdot \varepsilon_{act}, 0], 1 \leq k \leq n+1. \tag{4.4}$$

Then x^ν is ε_{stat}-stationary for SIP with $(2\varepsilon_{act})$-active indices $y^{k,\nu}$ and multipliers $(\kappa^\nu, \lambda^\nu)$, in contradiction to the assumption that the termination criterion is not satisfied for any $\nu \in \mathbb{N}$.

By the definition and by the overestimation property of g^k, (4.3) implies (4.4) if for all $k \in \{1, \ldots, n+1\}$ we can show

$$\lim_{\nu \to \infty} \frac{\alpha_k^\nu}{2}(y^{k,\nu} - \eta^{k-1,\nu})(\eta^{k,\nu} - y^{k,\nu}) = 0. \tag{4.5}$$

Because of

$$0 \leq \frac{\alpha_k^\nu}{2}(y^{k,\nu} - \eta^{k-1,\nu})(\eta^{k,\nu} - y^{k,\nu}) \leq \frac{\bar{\alpha}}{8}(\eta^{k,\nu} - \eta^{k-1,\nu})^2$$

(compare (2.6)) it is sufficient to show that the lengths $(\eta^{k,\nu} - \eta^{k-1,\nu})$ of $Y^{k,\nu}$ tend to zero for each $k \in \{1, \dots, n+1\}$.

In fact, Lemma 4.5 guarantees that in each iteration of Algorithm 3.4 at least one of the existing subdivision nodes $Y^{\nu,k}$ is divided into two subintervals with lengths bounded above by $(1 - \varepsilon_\cup)(\eta^{k,\nu} - \eta^{k-1,\nu})$. So for each $p \in \mathbb{N}$ at least one node with length bounded by $(1 - \varepsilon_\cup)^p$ is generated. Moreover, no subdivision node can be visited twice by the algorithm, so that for each $k \in \{1, \dots, n+1\}$ all $Y^{\nu,k}$, $\nu \in \mathbb{N}$, are pairwise different. As for each $p \in \mathbb{N}$ only finitely many nodes of length greater than $(1 - \varepsilon_\cup)^p$ exist, the lengths of $Y^{\nu,k}$, $\nu \in \mathbb{N}$, must tend to zero. ∎

5. Numerical examples. For the numerical illustrations in this section we implemented Algorithm 3.4 in *Matlab* 6.5 and used the routine *fmincon* from its *Optimization Toolbox* 2.2 with default tolerances as the black box NLP solver in Step 2 of the algorithm. In particular, the tolerance for constraint violations was 10^{-6}. All examples were run on a 1.2 GHz Pentium III processor.

For the examples in this article the convexification parameters can be determined analytically. We emphasize that, for more general examples which go beyond the scope of this article, we obtained good results using the *Matlab* toolbox *Intlab* 5.3 [41].

5.1. Chebyshev approximation. As a first test example we consider Chebyshev approximation of the function $\sin(\pi y)$ by a quadratic function on the interval $Y = [0, 1]$; that is, we wish to solve

$$\text{CA}: \qquad \min_{x \in \mathbb{R}^3} \| \sin(\pi y) - (x_3 y^2 + x_2 y + x_1) \|_{\infty,[0,1]},$$

with

$$\| \sin(\pi y) - (x_3 y^2 + x_2 y + x_1) \|_{\infty,[0,1]} = \max_{y \in [0,1]} | \sin(\pi y) - (x_3 y^2 + x_2 y + x_1) |.$$

For a survey on semi-infinite treatment of Chebyshev approximation problems see [23]. The epigraph reformulation of CA yields the equivalent problem

$$\min_{x \in \mathbb{R}^4} x_4 \quad \text{subject to} \quad \max_{y \in [0,1]} | \sin(\pi y) - (x_3 y^2 + x_2 y + x_1) | \le x_4$$

and thus the semi-infinite problem

$$\text{SIP}_{CA}: \qquad \min_{x \in \mathbb{R}^4} x_4 \quad \text{subject to} \quad \pm \left(\sin(\pi y) - x_3 y^2 - x_2 y - x_1 \right) \le x_4 \quad \text{for all } y \in [0, 1].$$

We search for a solution in the box $X = [x^\ell, x^u]$, with $x^\ell = (-1, 3, -5, -1)^\top$ and $x^u = (1, 5, -3, 3)^\top$.

Convexification parameters for the two semi-infinite constraints are readily available. In fact, for

$$g_\pm(x, y) = \pm \left(\sin(\pi y) - x_3 y^2 - x_2 y - x_1 \right) - x_4$$

we obtain $\nabla_y^2 g_\pm(x, y) = \mp \left(\pi^2 \sin(\pi y) + 2x_3 \right)$, so that an upper bound on $X \times Y$ is $\alpha_+ = 10$ for $\nabla_y^2 g_+$ and $\alpha_- = \pi^2 - 6$ for $\nabla_y^2 g_-$. Moreover, upper bounds on subdomains $[\eta^\ell, \eta^u] \subset [0, 1]$ are given by

$$\alpha_+(\eta^\ell, \eta^u) = \max \left(0, -\pi^2 \min(\sin(\pi \eta^\ell), \sin(\pi \eta^u)) + 10 \right), \tag{5.1}$$

$$\alpha_-(\eta^\ell, \eta^u) = \max \left(0, \pi^2 \theta(\eta^\ell, \eta^u) - 6 \right), \tag{5.2}$$

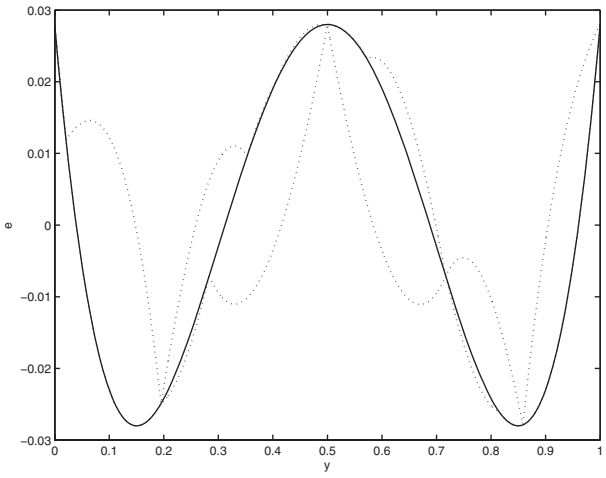


FIG. 1. *The optimal error function for SIP$_{CA}$ with convex relaxations.*

with

$$\theta(\eta^\ell, \eta^u) = \begin{cases} \sin(\pi\eta^u), & \eta^u \le 0.5, \\ 1, & \eta^\ell < 0.5 < \eta^u, \\ \sin(\pi\eta^\ell), & 0.5 \le \eta^\ell. \end{cases}$$

We initialize phase I with the (even infeasible) point $x^0 = (1, 1, 1, 1)^\top$ and use the tolerances $\varepsilon_{stat} = \varepsilon_{act} = 10^{-3}$ as well as $\varepsilon_\cup = 2\varepsilon_{act} \cdot \min(1, 1/\alpha_+, 1/\alpha_-)$.

In the following we compare the performance of Algorithm 3.4 for the two cases that either the uniform bounds $\alpha_\pm$ are used on all computed subdomains of Y or that $\alpha_\pm$ is adaptively improved on the subdomains according to (5.1), (5.2).

Case 1 (uniform convexification parameters). After one iteration (0.371 CPU seconds), phase I finds a feasible starting point $\bar{x}$ for SIP$_{CA}$ with the objective value $\bar{x}_4 = 1.54$. Note that, due to the *feasibility* of $\bar{x}$, this value is already a *valid upper bound* for the maximal error.

After nine more iterations (46.285 CPU seconds) the algorithm terminates with the point $x^\star = (-0.028, 4, -4, 0.028)^\top$ and objective value $x_4^\star = 0.028364$. The norm of the stationarity condition at $x^\star$ is less than 10^{-14}, and the solution of the problem P_{outer} with the obtained subdivision points E as discretization points yields the optimal value 0.028003. Since P_{outer} is a linear problem, this value is actually the *global* optimal value of P_{outer} and yields a valid *lower* bound for the optimal value of SIP$_{CA}$. As $x_4^\star = 0.028364$ is a valid *upper* bound for the latter optimal value, we have a certificate that we determined the globally optimal value of SIP$_{CA}$ within a tolerance of 10^{-3}.

Case 2 (adaptive convexification parameters). The performance of phase I is identical to Case 1, and after six more iterations (6.26 CPU seconds) the algorithm terminates with $x^\star = (-0.028, 4, -4, 0.028)^\top$ and objective value $x_4^\star = 0.028011$. Figure 1 illustrates the error function $e(x^\star, y) = \sin(\pi y) - (x_3^\star y^2 + x_2^\star y + x_1^\star)$ in the solution point, together with the convex relaxations upon termination. Note that the approximations of $e(x^\star, y)$ are depicted for both semi-infinite constraints, where in the figure one approximation appears as piecewise concave and the other as piecewise convex.

The norm of the stationarity condition at $x^\star$ is less than 10^{-9}, and the solution of the problem P_{outer} yields the optimal value 0.028004. We thus obtain a certificate that we determined the global optimal value of SIP_{CA} within a tolerance of 10^{-5}.

Comparison of the two cases shows that the use of adaptive convexification parameters can speed up the algorithm significantly, here almost by the factor 10 in CPU time. We emphasize, however, that in this example the formula for adaptive convexification parameters is explicitly available and easy to evaluate, whereas in other applications the additional computations by, for instance, interval methods can lead to a significant increase of CPU time.

5.2. Design centering. A design-centering problem considers a container set $C \subset \mathbb{R}^m$ and a parameterized body $B(x) \subset \mathbb{R}^m$, with parameter vector $x \in \mathbb{R}^n$. The task is to inscribe $B(x)$ into C such that some functional f (e.g., the volume of $B(x)$) is maximized:

$$\text{DC}: \quad \max_{x\in\mathbb{R}^n} f(x) \quad \text{subject to} \quad B(x) \subset C.$$

Design-centering problems with special sets $B(x)$ and C have been studied extensively; see, for instance, [18] for the complexity of inscribing a convex body into a convex container, [24] for maximization of a production yield under uncertain quality parameters, and [34, 50] for the problem of cutting a diamond with prescribed shape features and maximal volume from a raw diamond. Also cutting stock problems [8] and set containment problems [31] are examples of design centering. An implementable solution method is given in [35], and the so-called maneuverability problem of a robot from [17] is the first design-centering problem that was analyzed in its reformulation as a semi-infinite problem [22]. A structural analysis of the special semi-infinite problems arising as reformulated design-centering problems is given in [46].

Here we consider DC with a simply connected container set $C \subset \mathbb{R}^2$ and the parameterized body $B(x) = \{z \in \mathbb{R}^2 | \|z-(x_1, x_2)\|_2 \le x_3\}$, that is, a disk with variable midpoint (x_1, x_2) and radius x_3. We put $f(x) = x_3$, corresponding to maximization of the area of $B(x)$. Since C is simply connected, the constraint $B(x) \subset C$ is equivalent to $\partial B(x) \subset C$, where $\partial B(x) = \{z \in \mathbb{R}^2 | \|z - (x_1, x_2)\|_2 = x_3\}$ denotes the boundary of the disk. We can parameterize this set by $\partial B(x) = \{z(x, y) | y \in [0, 2\pi]\}$, with

$$z(x, y) = \left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + x_3 \begin{pmatrix} \cos(y) \\ \sin(y) \end{pmatrix} \right).$$

For the semi-infinite reformulation of DC we assume that the container set is described by finitely many smooth inequality constraints:

$$C = \{z \in \mathbb{R}^2 | c_i(z) \le 0, i \in I\}.$$

The constraint $B(x) \subset C$ is then equivalent to

$$c_i(z) \le 0 \quad \text{for all } z \in \partial B(x), i \in I,$$

and also to

$$g_i(x, y) := c_i(z(x, y)) \le 0 \quad \text{for all } y \in Y := [0, 2\pi], i \in I,$$

so that we arrive at the semi-infinite problem

$$\text{SIP}_{DC}: \qquad \max_{x\in\mathbb{R}^3} x_3 \quad \text{subject to} \quad g_i(x, y) \le 0 \quad \text{for all } y \in Y, i \in I.$$

If C is contained in the box X_C with side lengths ℓ_1, ℓ_2, then each feasible point x is necessarily contained in $X = X_C \times [0, \min(\ell_1, \ell_2)/2]$. For the computation of convexification parameters we calculate

$$\nabla_y^2 g_i(x,y) = x_3^2 \begin{pmatrix} -\sin(y) \\ \cos(y) \end{pmatrix}^\top \nabla^2 c_i(z(x,y)) \begin{pmatrix} -\sin(y) \\ \cos(y) \end{pmatrix} - x_3 \nabla^\top c_i(z(x,y)) \begin{pmatrix} \cos(y) \\ \sin(y) \end{pmatrix}$$

for each $i \in I$. The vectors with trigonometric entries have length one, so that we obtain coarse upper bounds for $\nabla_y^2 g_i$ as follows:

$$\begin{pmatrix} -\sin(y) \\ \cos(y) \end{pmatrix}^\top \nabla^2 c_i(z(x,y)) \begin{pmatrix} -\sin(y) \\ \cos(y) \end{pmatrix} \le \lambda_{max}\left(\nabla^2 c_i(z(x,y))\right),$$

with λ_{max} denoting the maximal eigenvalue of a symmetric matrix, and due to $x_3 \ge 0$

$$-x_3 \nabla^\top c_i(z(x,y)) \begin{pmatrix} \cos(y) \\ \sin(y) \end{pmatrix} \le x_3 \|\nabla c_i(z(x,y))\|_2.$$

A combination of these estimates yields for each $i \in I$

$$\begin{aligned} &\max_{(x,y)\in X\times Y} \nabla^2 g_i(x,y) \\ &\le \max_{(x,y)\in X\times Y} \left(x_3^2 \lambda_{max}\left(\nabla^2 c_i(z(x,y))\right) + x_3 \|\nabla c_i(z(x,y))\|_2\right). \end{aligned} \tag{5.3}$$

For our computational experiment we consider the functions

$$\begin{aligned} c_1(z) &= 0.3\sin(\pi z_1) - z_2, \\ c_2(z) &= z_1^2 + 0.3 z_2^2 - 1. \end{aligned}$$

The set C is then contained in $X_C = [-1.5, 1.5] \times [-1, 2]$, so that we can put $X = X_C \times [0, 1.5]$. With the aid of (5.3) we obtain uniform convexification parameters $\alpha_1 = 8.723$ and $\alpha_2 = 17.485$ for the constraints g_1 and g_2 , respectively. As the initial point x^0 we choose the midpoint of X, and we use the tolerances $\varepsilon_{stat} = \varepsilon_{act} = 10^{-3}$ as well as $\varepsilon_\cup = 2\varepsilon_{act} \cdot \min(1, 1/\alpha_1, 1/\alpha_2)$.

After 15 iterations (22.793 CPU seconds) phase I finds a feasible starting point for SIP_{DC}, and the corresponding approximating problem has the solution $\bar{x} = (-0.116, 1.026, 0.397)^\top$ with the objective value $\bar{x}_3 = 0.397$. Due to the *feasibility* of $\bar{x}$, the corresponding disk $B(\bar{x})$ is guaranteed to be contained in C. Figure 2 illustrates this fact.

After 57 more iterations (181.566 CPU seconds) the algorithm terminates with the point $x^\star = (0, 0.962, 0.776)^\top$ with objective value $x_3^\star = 0.77612$, illustrated in Figure 3.

The norm of the stationarity condition at $x^\star$ is less than 10^{-8}, and the solution of the problem P_{outer} yields the optimal value 0.77697. However, since P_{outer} is a nonconvex problem, this value is *not* necessarily an upper bound for the optimal value of SIP_{DC}. Thus, we do not obtain a certificate for global optimality in this example.

While Algorithm 3.4 can certainly be accelerated by exploiting special features of SIP_{DC} (see also section 6), for certain applications a user might have to terminate it prematurely due to time restrictions. It is worth pointing out that even then the last iterate is feasible; that is, at least *some* disk $B(\bar{x}) \subset C$ is generated.

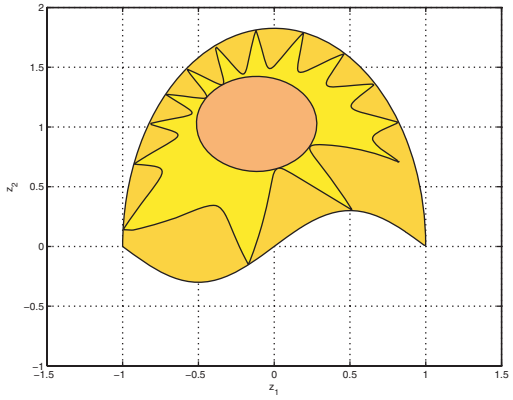


FIG. 2. *A container C with approximating sets and an inscribed disk.*

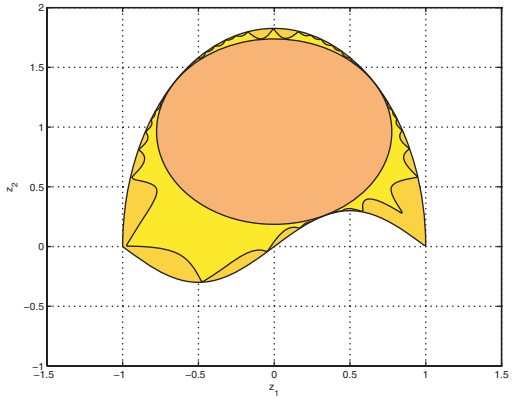


FIG. 3. *Design-centering solution of Algorithm* 3.4.

6. Future challenges.

6.1. Improvements. In this article we presented the adaptive convexification algorithm in its simplest form. Although it can be improved in a number of ways, our aim was to explain its basic ideas and to show that these already result in a well-defined, convergent, and implementable method. Convergence proofs for the following possible improvements are subject of future research.

A major potential for acceleration of the algorithm certainly lies in the magnitude of the convexification parameters and in their adaptive refinement. Lemma 3.2(iii) and Proposition 4.1 quantify the obvious fact that tighter convex relaxations, that is, better bounds in (3.3), (3.4), can speed up the algorithm. Moreover, tighter relaxations also lead to a less restrictive bound on the tolerance $\varepsilon_\cup$ in Step 1 of Algorithm 3.4.

As mentioned in section 2.3, tighter relaxations are already available for a large number of specially structured functions [12]. Improvements of the αBB method itself like the generalized αBB method [3] also lead to significantly tighter convex relaxations, however, at higher computational cost. In special examples also modeling of the problem can influence the size of convexification parameters. In design centering they depend, for instance, on the scaling of the functions c_i, $i \in I$.

We emphasize that in (3.3) we only consider adaptation of the convexification parameters with respect to the index variable y while they are uniform in the decision variable $x \in X$. Additional adaptation with respect to x is possible by restricting the distance of a new iterate $x^{\nu+1}$ generated by Step 2 of Algorithm 3.4 to a ball $B^\nu \subset X$ around the current iterate x^ν. Tighter relaxations can then be computed using (3.3) with X replaced by B^ν. A main problem with this approach is that convexification parameters of previous steps may become invalid during the iteration, even if they correspond to inactive constraints. Hence, the relaxations on all subdomains have to be recomputed in each step, giving rise to a possibly high computational effort. To alleviate this difficulty, ideas from trust region methods might help.

As we pointed out, the number of new constraints in each step of Algorithm 3.4 is linear in n, as opposed to exponential growth for the exhaustive subdivision from Proposition 4.1. A possibility to further diminish the number of added constraints in each step, in the spirit of exchange methods, is to add only those active approximating constraints with the largest benefit for the activity of the original constraint; that is, $y \in Y_0^{\alpha BB}(\bar{x})$ only enters the $\varepsilon_\cup$-union with E if y minimizes $g(\bar{x}, y)$ over $Y_0^{\alpha BB}(\bar{x})$.

6.2. Generalizations. Algorithm 3.4 can be generalized straightforwardly to SIP with finitely many semi-infinite constraints (i.e., $p \geq 1$) and finitely many additional smooth equality and inequality constraints. Our implementation for the solution of the problems in section 5 actually handles this situation.

The generalization to higher dimensional box index sets Y (i.e., $m \geq 1$) is merely of technical nature. The two main new tasks are to manage the bookkeeping for the adaptive subdivisions of Y and to compute convexification parameters in higher dimensions. The latter problem can be solved with the techniques of αBB in higher dimensions [1, 2, 12]. There the main idea is to add a separable quadratic relaxation function to g in (3.2), with one parameter for each component of the vector y. Sufficiently large parameters then again lead to concavity, where, for instance, a Gerschgorin approach can guarantee the negative semi-definiteness of the interval Hessian matrix on a subdomain [1, 2, 12].

If the index set Y is not given in box form and cannot be simply transformed into a box, a possible approach is to approximate Y appropriately by boxes. Here again the trivial observation comes into play that a semi-infinite constraint can be equivalently replaced by finitely many semi-infinite constraints corresponding to a subdivision of the index set. A more efficient possibility might be to construct triangulations of Y, but then the αBB ideas also have to be transferred to triangular domains.

6.3. Open questions. In addition to outstanding convergence proofs for the above improvements and the implementation issues for higher dimensional index sets, our approach leads to a number of further open questions.

Regarding the stationarity concept, if the natural assumption EMFCQ [26] is satisfied everywhere in M, Algorithm 3.4 with the theoretical tolerances $\varepsilon_{stat} = \varepsilon_{act} = 0$ would of course generate a Karush–Kuhn–Tucker point of SIP. We conjecture that under EMFCQ this is also the case for sufficiently small tolerances $\varepsilon_{stat}, \varepsilon_{act} > 0$.

We focused our attention on the convergence of stationary points of approximating problems $\mathrm{SIP}_{\alpha BB}(E, \alpha)$ to a stationary point of SIP. Since we can actually expect the black box NLP solver to find local minimizers of $\mathrm{SIP}_{\alpha BB}(E, \alpha)$, one can try to prove convergence of such points to a local minimizer of SIP. At least in our numerical experiments we observed this convergence throughout.

The main purpose of Algorithm 3.4 is the generation of feasible iterates for SIP. Under weak assumptions these iterates should actually be *interior points* of the original feasible set M. We observed this in the numerical experiments, too.

Although our numerical results seem very promising, further work is needed on error estimates on the numerical solution of the auxiliary problem $P(E, \alpha)$. In fact, feasibility of the iterates may be jeopardized if feasibility and local optimality for $P(E, \alpha)$ are guaranteed only up to some tolerances. Handling this issue is beyond the scope of the present article, which is meant to give the basic ideas of adaptive convexification.

Regarding the performance of the algorithm, in sections 5.1 and 6.1 we have seen that adaptation of the convexification parameters as well as more sophisticated techniques for their computation can result in a significant acceleration of computing time. There is, however, an obvious tradeoff with the additional computational effort of these modifications. General statements about this tradeoff would be helpful for applications.

Acknowledgments. We thank the anonymous referees for their precise and substantial remarks which led to a significantly improved version of this article. The second author thanks the Department of Chemical Engineering at Princeton University for their hospitality during his visit, whereupon this work originated.

REFERENCES

[1] C. S. Adjiman, I. P. Androulakis, and C. A. Floudas, *A global optimization method, αBB, for general twice-differentiable constrained NLPs - I: Theoretical advances*, Computers and Chemical Engineering, 22 (1998), pp. 1137–1158.

[2] C. S. Adjiman, I. P. Androulakis, and C. A. Floudas, *A global optimization method, αBB, for general twice-differentiable constrained NLPs - II: Implementation and computational results*, Computers and Chemical Engineering, 22 (1998), pp. 1159–1179.

[3] I. G. Akrotirianakis and C. A. Floudas, *A new class of improved convex underestimators for twice continuously differentiable constrained NLPs*, J. Global Optim., 30 (2004), pp. 367–390.

[4] B. Bhattacharjee, W. H. Green, Jr., and P. I. Barton, *Interval methods for semi-infinite programs*, Comput. Optim. Appl., 30 (2005), pp. 63–93.

[5] B. Bhattacharjee, P. Lemonidis, W. H. Green, Jr., and P. I. Barton, *Global solution of semi-infinite programs*, Math. Program., 103 (2005), pp. 283–307.

[6] B. Brosowski, *Parametric Semi-infinite Optimization*, Peter Lang, Frankfurt, 1982.

[7] B. Chen and P. T. Harker, *A non-interior-point continuation method for linear complementarity problems*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 1168–1190.

[8] V. Chvatal, *Linear Programming*, Freeman, New York, 1983.

[9] J. M. Danskin, *The Theory of Max-Min and its Applications to Weapons Allocation Problems*, Springer, New York, 1967.

[10] A. Fischer, *A special Newton-type optimization method*, Optimization, 24 (1992), pp. 269–284.

[11] R. Fletcher, S. Leyffer, D. Ralph, and S. Scholtes, *Local convergence of SQP methods for mathematical programs with equilibrium constraints*, SIAM J. Optim., 17 (2006), pp. 259–286.

[12] C. A. Floudas, *Deterministic Global Optimization, Theory, Methods and Applications*, Kluwer, Dordrecht, 2000.

[13] C. A. Floudas, *Research challenges, opportunities and synergism in systems engineering and computational biology*, AIChE J., 51 (2005), pp. 1872–1884.

[14] C. A. Floudas, I. G. Akrotirianakis, S. Caratzoulas, C. A. Meyer, and J. Kallrath, *Global optimization in the 21st century: Advances and challenges*, Computers and Chemical Engineering, 29 (2005), pp. 1185–1202.

[15] K. Glashoff and S. A. Gustafson, *Linear Optimization and Approximation*, Springer, Berlin, 1983.

[16] M. A. Goberna and M. A. López, *Linear Semi-infinite Optimization*, Wiley, Chichester, 1998.

[17] T. J. Graettinger and B. H. Krogh, *The acceleration radius: A global performance measure for robotic manipulators*, IEEE Journal of Robotics and Automation, 4 (1988), pp. 60–69.
[18] P. Gritzmann and V. Klee, *On the complexity of some basic problems in computational convexity.* I. *Containment problems*, Discrete Math., 136 (1994), pp. 129–174.
[19] E. Hansen, *Global Optimization using Interval Analysis*, Marcel Dekker, New York, 1992.
[20] R. Hettich and H. Th. Jongen, *Semi-infinite programming: Conditions of optimality and applications*, in Optimization Techniques, Part 2, J. Stoer, ed., Lecture Notes in Control and Information Sciences 7, Springer, Berlin, 1978, pp. 1–11.
[21] R. Hettich and K. O. Kortanek, *Semi-infinite programming: Theory, methods, and applications*, SIAM Rev., 35 (1993), pp. 380–429.
[22] R. Hettich and G. Still, *Semi-infinite programming models in robotics*, in Parametric Optimization and Related Topics II, J. Guddat, H. Th. Jongen, B. Kummer, and F. Nožička, eds., Akademie Verlag, Berlin, 1991, pp. 112–118.
[23] R. Hettich and P. Zencke, *Numerische Methoden der Approximation und Semi-Infiniten Optimierung*, Teubner, Stuttgart, 1982.
[24] R. Horst and H. Tuy, *Global Optimization*, Springer, Berlin, 1990.
[25] F. John, *Extremum problems with inequalities as subsidiary conditions*, in Studies and Essays, R. Courant Anniversary Volume, Interscience, New York, 1948, pp. 187–204.
[26] H. Th. Jongen, F. Twilt, and G.-W. Weber, *Semi-infinite optimization: Structure and stability of the feasible set*, J. Optim. Theory Appl., 72 (1992), pp. 529–552.
[27] H. Th. Jongen and G. W. Weber, *Nonlinear optimization: Characterization of structural stability*, J. Global Optim., 1 (1991), pp. 47–64.
[28] C. Kanzow, *Some noninterior continuation methods for linear complementarity problems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 851–868.
[29] M. Kočvara, J. Outrata, and J. Zowe, *Nonsmooth Approach to Optimization Problems with Equilibrium Constraints: Theory, Applications and Numerical Results*, Kluwer, Dordrecht, 1998.
[30] Z. Luo, J. Pang, and D. Ralph, *Mathematical Programs with Equilibrium Constraints*, Cambridge University Press, Cambridge, 1996.
[31] O. L. Mangasarian, *Set containment characterization*, J. Global Optim., 24 (2002), pp. 473–480.
[32] C. D. Maranas and C. A. Floudas, *Global minimum potential energy conformations for small molecules*, J. Global Optim., 4 (1994), pp. 135–170.
[33] A. Neumaier, *Interval Methods for Systems of Equations*, Cambridge University Press, Cambridge, 1990.
[34] V. H. Nguyen and J. J. Strodiot, *Computing a global optimal solution to a design centering problem*, Math. Program., 53 (1992), pp. 111–123.
[35] E. Polak, *An implementable algorithm for the optimal design centering, tolerancing and tuning problem*, J. Optim. Theory Appl., 37 (1982), pp. 45–67.
[36] E. Polak, *On the mathematical foundation of nondifferentiable optimization in engineering design*, SIAM Rev., 29 (1987), pp. 21–89.
[37] E. Polak, *Optimization. Algorithms and Consistent Approximations*, Springer, Berlin, 1997.
[38] R. Reemtsen and S. Görner, *Numerical methods for semi-infinite programming: A survey*, in Semi-Infinite Programming, R. Reemtsen and J.-J. Rückmann, eds., Kluwer, Boston, 1998, pp. 195–275.
[39] S. M. Robinson, *Stability theory for systems of inequalities, part* II: *Differentiable nonlinear systems*, SIAM J. Numer. Anal., 13 (1976), pp. 497–513.
[40] J.-J. Rückmann and O. Stein, *On convex lower level problems in generalized semi-infinite optimization*, in Semi-infinite Programming—Recent Advances, M. A. Goberna and M. A. López, eds., Kluwer, Dordrecht, 2001, pp. 121–134.
[41] S. M. Rump, *INTLAB - INTerval LABoratory*, Institute for Reliable Computing, Hamburg University of Technology, 1999, http://www.ti3.tu-harburg.de/rump/intlab.
[42] H. Scheel and S. Scholtes, *Mathematical programs with complementarity constraints: Stationarity, optimality, and sensitivity*, Math. Oper. Res., 25 (2000), pp. 1–22.
[43] S. Scholtes and M. Stöhr, *Exact penalization of mathematical programs with equilibrium constraints*, SIAM J. Control Optim., 37 (1999), pp. 617–652.
[44] S. Smale, *Algorithms for solving equations*, in Proceedings of the International Congress of Mathematicians, AMS, Providence, 1987, pp. 172–195.
[45] O. Stein, *Bi-level Strategies in Semi-infinite Programming*, Kluwer, Boston, 2003.
[46] O. Stein, *A semi-infinite approach to design centering*, in Optimization with Multivalued Mappings: Theory, Applications and Algorithms, S. Dempe and S. Kalashnikow, eds., Springer, New York, 2006, pp. 209–228.

[47] O. Stein and G. Still, *On generalized semi-infinite optimization and bilevel optimization*, European J. Oper. Res., 142 (2002), pp. 444–462.
[48] O. Stein and G. Still, *Solving semi-infinite optimization problems with interior point techniques*, SIAM J. Control Optim., 42 (2003), pp. 769–788.
[49] W. Wetterling, *Definitheitsbedingungen für relative extrema bei optimierungs- und approximationsaufgaben*, Numer. Math., 15 (1970), pp. 122–136.
[50] A. Winterfeld, *Maximizing volumes of lapidaries by use of hierarchical GSIP-models*, Diploma thesis, Technische Universität Kaiserslautern and Fraunhofer Institut für Techno- und Wirtschaftsmathematik, 2004.

SIAM J. OPTIM.
Vol. 18, No. 4, pp. 1209–1230

ON THE CONVERGENCE OF AUGMENTED LAGRANGIAN METHODS FOR CONSTRAINED GLOBAL OPTIMIZATION*

H. Z. LUO†, X. L. SUN‡, AND D. LI§

Abstract. In this paper, we present new convergence properties of the primal-dual method based on four types of augmented Lagrangian functions in the context of constrained *global* optimization. Convergence to a global optimal solution is first established for a basic primal-dual scheme under standard conditions. We then prove this convergence property for a modified augmented Lagrangian method using a safeguarding strategy without appealing to the boundedness assumption of the multiplier sequence. We further show that, under the same weaker conditions, the convergence to a global optimal solution can still be achieved by either modifying the multiplier updating rule or normalizing the multipliers in augmented Lagrangian methods.

Key words. nonconvex optimization, constrained global optimization, augmented Lagrangian functions, modified augmented Lagrangian methods, convergence to global solution

AMS subject classifications. 90C26, 90C46

DOI. 10.1137/060667086

1. Introduction.

We consider in this paper the following inequality-constrained nonlinear optimization problem:

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

where f and all g_i, $i = 1, \ldots, m$, are continuously differentiable functions and X is a nonempty closed set in $\mathbb{R}^n$. Note that f and g_i, $i = 1, \ldots, m$, are not necessarily convex functions.

Lagrangian dual methods have been serving as a fundamental solution methodology in convex programming. It is well known, however, that classical Lagrangian dual methods may fail to identify the optimal solution of the nonconvex problem (P) due to the existence of a duality gap. Augmented Lagrangians have been proposed as a remedy to alleviate this situation. The first augmented Lagrangian method was independently proposed by Hestenes [13] and Powell [39] for equality-constrained problems by introducing a quadratic penalty term in the Lagrangian function. This method was extended by Rockafellar [40], [41] to deal with inequality constraints. Rockafellar's augmented Lagrangian function for (P) is continuously differentiable with respect to x but not twice differentiable even when f and all g_i are twice differentiable functions, thus preventing the use of Newton-type methods from solving the corresponding unconstrained Lagrangian relaxation problems. Mangasarian's augmented Lagrangian

*Received by the editors August 8, 2006; accepted for publication (in revised form) June 7, 2007; published electronically October 10, 2007. This work was supported by the National Natural Science Foundation of China grants 70671064 and 60473097 and the Research Grants Council of Hong Kong grant CUHK 4245/04E.

http://www.siam.org/journals/siopt/18-4/66708.html

†Department of Applied Mathematics, Zhejiang University of Technology, Hangzhou, Zhejiang 310032, People's Republic of China (hzluo@zjut.edu.cn).

‡Department of Management Science, School of Management, Fudan University, Shanghai 200433, People's Republic of China (xls@fudan.edu.cn).

§Corresponding author. Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, N. T., Hong Kong (dli@se.cuhk.edu.hk).

[28], the exponential-type penalty function [6], [20], and the modified barrier function [35] are three important classes of twice differentiable augmented Lagrangian functions. Various augmented Lagrangians or nonlinear Lagrangians have also been proposed, and their duality and exact penalization properties have been studied (see, e.g., [16], [23], [24], [31], [32], [33], [44], [45], [52]).

Local convergence properties of augmented Lagrangian methods for nonconvex optimization have been studied by many researchers. Mangasarian [28] analyzed the local convergence of a class of augmented Lagrangians that include Rockafellar's augmented Lagrangian as a special case. Nguyen and Strodiot [30] proved the local convergence of the modified exponential Lagrangian method. The local convergence of the modified barrier Lagrangian method and log-sigmoid multiplier method for nonconvex problems was analyzed in [35] and [36], respectively. Contesse-Becker [**?**] proved the local convergence of Rockafellar's augmented Lagrangian method without the strict complementarity condition. Huang and Yang [17] showed that the first-order and second-order necessary optimality conditions of Rockafellar's augmented Lagrangian problems converge to those of the original problem. Polak and Tits [34], Hager [12], Bartholomew-Biggs [3], and Yamashita [53] investigated the global convergence of Rockafellar's augmented Lagrangian methods for nonconvex inequality-constrained problems. The convexification effect of adopting pth-power Lagrangian functions has been revealed for augmented Lagrangian functions in [25] and for the perturbation function in [23], [24]. The existence of a global saddle point of certain augmented or nonlinear Lagrangian functions has been investigated in [26], [27], [47].

Global convergence of the augmented Lagrangian method for nonconvex equality-constrained problems was analyzed in [6], [39]. An indispensable assumption in most existing global convergence analyses for augmented Lagrangian methods is that the multiplier sequence generated in the algorithms is bounded. This restrictive assumption confines applications of augmented Lagrangian methods in many situations. Conn et al. [8], Conn, Gould, and Toint [9], and Lewis and Torczon [22] presented modified augmented Lagrangian methods for nonconvex optimization with equality constraints and proved global convergence results without appealing to this assumption. Andreani et al. [1], [2] and Birgin, Castillo, and Martínez [7] investigated the augmented Lagrangian methods using safeguarding strategies for nonconvex constrained problems. Global convergence was established in [1], [2], [7] without requiring the boundedness condition of the multiplier sequence. Global convergence of augmented Lagrangian methods for convex programming has also been studied in [4], [6], [19], [21], [40], [42], [50]. To our knowledge, the convergence of augmented Lagrangian methods to a global solution of a constrained nonconvex global optimization problem has been investigated only in [41] for Rockafellar's augmented Lagrangian method and in [6] for the multiplier method of Hestenes and Powell for equality-constrained problems.

Constrained global optimization has been one of the challenging subjects in nonlinear optimization. On one hand, implementable methods for constrained global optimization have been developed only for some special problems such as concave minimization (see [5], [43]) and monotone optimization (see [48], [51]). On the other hand, various deterministic and stochastic methods have been proposed for unconstrained global optimization (see, e.g., [14], [15]). It is therefore desirable to reduce a constrained global optimization problem into a sequence of unconstrained global optimization problems so that the methods developed for unconstrained global optimization can be used. This solution scheme was adopted in constructing auxiliary functions for constrained global optimization (see [46], [49]).

The focus of this paper is on the convergence properties of augmented Lagrangian methods in the context of constrained global optimization. We will study four classes of prominent augmented Lagrangians in the literature: (i) Kort and Bertsekas's augmented Lagrangian, (ii) Mangasarian's augmented Lagrangian, (iii) the exponential-type augmented Lagrangian, and (iv) the modified barrier function. As we will see later in the paper, these four classes of augmented Lagrangians include many existing important augmented Lagrangian functions as their special cases. We first show that, under standard conditions, the basic primal-dual method employing these four classes of augmented Lagrangian functions converges to a global solution of (P) if the unconstrained relaxation problems is solved globally. Three modified augmented Lagrangian methods are then proposed that adopt a safeguarding strategy, modify multiplier updating criteria, and normalize the multiplier, respectively. Convergence results are proved for these modified augmented Lagrangian methods without appealing to the boundedness of the multiplier sequence. The convergence results obtained in this paper provide theoretical foundations for applying augmented Lagrangian methods in constrained global optimization.

The paper is organized as follows. In section 2, we describe four classes of augmented Lagrangian functions. We present the convergence results for the basic primal-dual method in section 3. The modified primal-dual scheme with safeguarding is investigated in section 4. In section 5, we establish the convergence results for the modified primal-dual scheme with conditional multiplier updating. The normalized multiplier method is discussed in section 6. Finally, some concluding remarks are given in section 7.

2. Four classes of augmented Lagrangian functions. In this section, we describe four classes of augmented Lagrangian functions. The augmented Lagrangian methods studied in this paper are based on these four classes of functions.

The following general class of augmented Lagrangians for (P) was introduced by Kort and Bertsekas [6], [21]:

$$L_1(x,\lambda,p) = f(x) + \frac{1}{p}\sum_{i=1}^{m} W(pg_i(x),\lambda_i), \tag{2.1}$$

where $p > 0$, $x \in X$, $\lambda = (\lambda_1,\dots,\lambda_m)^T \geq 0$, and the function $W : \mathbb{R}^2 \to \mathbb{R}$ satisfies the following conditions:

(A1) W is continuously differentiable on $\mathbb{R}\times(0,\infty)$ and possesses, for all $s \in \mathbb{R}$, the right derivative

$$\lim_{t\to 0^+} \frac{W(s,t) - W(s,0)}{t};$$

(A2) $W(s,\cdot)$ is concave on $\mathbb{R}_+$ for each fixed $s \in \mathbb{R}$, where $\mathbb{R}_+ = [0,\infty)$;

(A3) for each fixed $t \in \mathbb{R}_+$, $W(\cdot,t)$ is convex on $\mathbb{R}$ and satisfies the following strict convexity condition: If $s_0 > 0$ or $W'_s(s_0,t) > 0$, then $W(s,t) - W(s_0,t) > (s-s_0)W'_s(s_0,t)$ for $s \neq s_0$, where $W'_s(s,t)$ denotes the partial derivative of $W(s,t)$ with respect to s;

(A4) $W(0,t) = 0$, $W'_s(0,t) = t$ for all $t \in \mathbb{R}_+$;

(A5) $\lim_{s\to-\infty} W'_s(s,t) = 0$ for all $t \in \mathbb{R}_+$;

(A6) $\inf_{s\in\mathbb{R}} W(s,t) > -\infty$ for all $t \in \mathbb{R}_+$;

(A7) $\lim_{s\to\infty} \frac{W(s,t)}{s} = \infty$ for all $t \in \mathbb{R}_+$.

By the convexity of $W(\cdot,t)$, condition (A7) is equivalent to $\lim_{s\to\infty} W'_s(s,t)=\infty$ for all $t\in\mathbb{R}_+$. It then follows from (A3) and (A5) that

$$W'_s(s,t)\geq 0 \quad \forall (s,t)\in\mathbb{R}\times\mathbb{R}_+. \tag{2.2}$$

Special cases of L_1 include the augmented function in [21], the modified Courant-type augmented Lagrangian function [6], [39], the penalized exponential-type augmented Lagrangian [6], and the augmented Lagrangian functions in [18], [29], [40], [41]. For example, the augmented function introduced in [6], [21] is a special case of $L_1(x,\lambda,p)$ with $W(s,t)$ defined as

$$W(s,t)=\begin{cases} ts+\phi(s) & \text{if } t+\phi'(s)\geq 0,\\ \min_{\tau\in\mathbb{R}}[t\tau+\phi(\tau)] & \text{otherwise,}\end{cases} \tag{2.3}$$

where $\phi:\ \mathbb{R}\to\mathbb{R}$ satisfies the following conditions:

(E1) $\phi(\cdot)$ is a strictly convex function and continuous differentiable on $\mathbb{R}$;

(E2) $\phi(0)=0$, $\phi'(0)=0$, $\lim_{s\to-\infty}\phi'(s)=-\infty$;

(E3) $\frac{\phi(s)}{|s|}\to\infty$, $(|s|\to\infty)$.

It can be verified that $W(s,t)$ defined in (2.3) satisfies (A1)–(A7) and that Rockafellar's augmented Lagrangian function is a special case of L_1 with $\phi(s)=(1/2)s^2$ in $W(s,t)$ defined in (2.3).

Mangasarian [28] proposed another general class of augmented Lagrangians for (P). Let

$$L_2(x,\lambda,p)=f(x)+\frac{1}{p}\sum_{i=1}^{m}\left[\theta(pg_i(x)+\lambda_i)_+-\theta(\lambda_i)\right],\quad x\in X,\ \lambda\in\mathbb{R}^m, \tag{2.4}$$

where $p>0$, $\theta(s)_+=\begin{cases}\theta(s), & s\geq 0\\ 0, & s<0\end{cases}$, and the function $\theta:\ \mathbb{R}\to\mathbb{R}$ satisfies the following conditions:

(B1) θ is continuously differentiable and strictly convex on $\mathbb{R}$;

(B2) $\theta(0)=0$, θ' maps $\mathbb{R}$ onto $\mathbb{R}$ and $\theta'(0)=0$;

(B3) $\frac{\theta(s)}{|s|}\to\infty$, $(|s|\to\infty)$.

Let $W(s,t)=\theta(s+t)_+-\theta(t)$. It can be verified that $W(s,t)$ satisfies conditions (A1)–(A7) except (A2) (see Remark 2.13 in [28]). Note that L_2 also includes Rockafellar's augmented Lagrangian as a special case by setting $\theta(s)=\frac{1}{2}s^2$ (see [40], [41]). We also see that if θ is twice differentiable and $\theta''(0)=0$, then $L_2(x,\lambda,p)$ is twice differentiable with respect to x when f and all g_i are twice differentiable. Examples of θ that satisfy conditions (B1)–(B3) and $\theta''(0)=0$ include $\theta(s)=\frac{1}{\rho}|s|^\rho$, where $\rho>2$, $\theta(s)=\frac{1}{2}(e^s+e^{-s})-\frac{1}{2}s^2-1$, and $\theta(s)=\frac{1}{2}\left[(e^s+e^{-s})/2-1\right]^2$.

It is easy to see that conditions (B1)–(B2) imply that θ' is a strictly increasing function on $\mathbb{R}$ and $\theta(s)\geq 0$ for $s\in\mathbb{R}$. Furthermore, θ is strictly decreasing on $(-\infty,0]$ and strictly increasing on $[0,\infty)$. By the convexity of θ, condition (B3) is equivalent to $\theta'(s)\to\infty$ $(s\to\infty)$.

The exponential-type augmented Lagrangian functions can be derived by applying classical Lagrangian formulation to a reformulation of (P). Let $\psi:\ \mathbb{R}\to\mathbb{R}$ be a function satisfying the following conditions:

(C1) ψ is a continuously differentiable and strictly convex function on $\mathbb{R}$;

(C2) $\psi(0)=0$, $\psi'(0)=1$, $\lim_{t\to-\infty}\psi'(t)=0$;

(C3) $\lim_{t\to-\infty}\psi(t)>-\infty$, $\lim_{t\to\infty}\frac{\psi(t)}{t}=\infty$.

From condition (C1), ψ' is strictly increasing on $\mathbb{R}$, which together with (C2) implies that $\psi'(t) > 0$ for any t. Thus ψ is a strictly increasing function on $\mathbb{R}$. An obvious example of ψ that satisfies (C1)–(C3) is $\psi(t) = \mathrm{e}^t - 1$. Other interesting examples of ψ include the hyperbolic-exponential function, the hyperbolic-quadratic function, the exponential-quadratic function (see [7], [11], [38]), and the modified log-sigmoid function [37].

Now problem (P) can be rewritten in an equivalent form:

$$\begin{aligned} &\min\ f(x)\\ &\text{s.t.}\ \ (1/p)\psi(pg_i(x)) \le 0,\ i = 1, \dots, m,\\ &\qquad x \in X, \end{aligned} \tag{2.5}$$

where $p > 0$ is a parameter. Applying the classical Lagrangian formulation to (2.5) leads to the following exponential-type augmented Lagrangian:

$$L_3(x, \lambda, p) = f(x) + \frac{1}{p}\sum_{i=1}^{m} \lambda_i \psi(pg_i(x)), \qquad x \in X,\ \lambda \ge 0. \tag{2.6}$$

To avoid ill-conditional numerical behavior of the exponential-type augmented Lagrangian function, a modified barrier function was proposed by Polyak [35]. Good numerical behavior of the modified barrier functions was shown in several papers (see, e.g., [4], [11]). We now consider a general class of modified barrier functions. Let φ: $(-\infty, 1) \to \mathbb{R}$ satisfy the following conditions:

(D1) φ is a continuously differentiable and strictly convex function on $(-\infty, 1)$;

(D2) $\varphi(0) = 0$, $\varphi'(0) = 1$, $\lim_{s\to-\infty} \varphi'(s) = 0$;

(D3) $\lim_{s\to-\infty} \frac{\varphi(s)}{s} = 0$.

Note that conditions (D1) and (D2) imply that φ is a strictly increasing function on $(-\infty, 1)$. Let $\Omega_p = \{x \in X \mid pg_i(x) < 1,\ i = 1, \dots, m\}$. Define the modified barrier augmented Lagrangian as follows:

$$L_4(x, \lambda, p) = \begin{cases} f(x) + \frac{1}{p}\sum_{i=1}^{m} \lambda_i \varphi(pg_i(x)), & x \in \Omega_p,\\ \infty, & x \in X \setminus \Omega_p, \end{cases} \tag{2.7}$$

where $p > 0$ and $\lambda \ge 0$. Taking $\varphi(s) = -\ln(1 - s)$ or $\varphi(s) = 1/(1 - s) - 1$ in (2.7) gives rise to the modified Frish function or the modified Carroll function introduced in [35], respectively.

3. Basic augmented Lagrangian method. In this section, we discuss the basic primal-dual method using the above four classes of augmented Lagrangians L_j $(j = 1, 2, 3, 4)$.

Define $h(x, \lambda, p) = (h_1(x, \lambda, p), \dots, h_m(x, \lambda, p))^T$, with

$$h_i(x, \lambda, p) = \begin{cases} (1/p)\left[W_s'\left(pg_i(x), \lambda_i\right) - \lambda_i\right] & \text{for } L_1,\\ (1/p)\left[\theta'\left(pg_i(x) + \lambda_i\right)_+ - \theta'(\lambda_i)\right] & \text{for } L_2,\\ (\lambda_i/p)\left[\psi'\left(pg_i(x)\right) - 1\right] & \text{for } L_3,\\ (\lambda_i/p)\left[\varphi'\left(pg_i(x)\right) - 1\right] & \text{for } L_4. \end{cases} \tag{3.1}$$

Algorithm 1 (basic primal-dual method).

Step 0 *(initialization). Select two positive sequences* $\{p_k\}_{k=0}^{\infty}$ *and* $\{\epsilon_k\}_{k=0}^{\infty}$, *with* $\epsilon_k \to 0$ *as* $k \to \infty$. *Choose* $\lambda^0 > 0$. *Set* $k = 0$.

Step 1 *(relaxation problem). Compute an* $x^k \in X$ *such that*

$$L_j(x^k, \lambda^k, p_k) \le \min_{x \in X} L_j(x, \lambda^k, p_k) + \epsilon_k \quad (j = 1, 2, 3, 4). \tag{3.2}$$

Step 2 *(multiplier updating). The multiplier vector* λ^k *is updated by the following formulas* ($i = 1, \ldots, m$)*:*

$$\lambda_i^{k+1} = \lambda_i^k + p_k h_i(x^k, \lambda^k, p_k). \tag{3.3}$$

Step 3. *Set* $k = k + 1$*; go to Step* 1.

Remark 1. For L_j $(j = 1, 3, 4)$, the multiplier updating formulas (3.3) are derived by noticing the following fact:

$$\nabla_x L_j(x^k, \lambda^k, p_k) = 0 \Rightarrow \nabla_x L(x^k, \lambda^{k+1}) = 0,$$

where L is the classical Lagrangian function given by $L(x, \lambda) = f(x) + \sum_{i=1}^m \lambda_i g_i(x)$. Note that $\frac{\partial L_2(x,\lambda,p)}{\partial \lambda_i} = \frac{1}{p}[\theta'(pg_i(x) + \lambda_i)_+ - \theta'(\lambda_i)]$. The multiplier updating formula (3.3) for L_2 can be viewed as executing a steepest ascent step for maximizing the function $L_2(x^k, \cdot, p_k)$ with step size p_k.

To ensure Step 1 is well posed, we need the following assumptions.

Assumption 1. $\underline{f} = \inf_{x \in X} f(x) > -\infty$.

Assumption 2. $\underline{g} = \inf_{x \in X} \min_{1 \le i \le m} g_i(x) > -\infty$.

Obviously, a sufficient condition to ensure Assumptions 1 and 2 is that X is a compact set. It can be verified that, under Assumption 1, Step 1 is well defined for L_j $(j = 1, 2, 3)$ and that Step 1 is well defined for L_4 under Assumptions 1 and 2.

Throughout the paper, we assume that Assumption 1 is always satisfied for L_j $(j = 1, 2, 3)$ and Assumptions 1 and 2 are always satisfied for L_4.

The following theorem can be viewed as a generalization of Proposition 2.1 in [6] and can be proved by using arguments similar to [6].

THEOREM 1. *Assume that* (A1)–(A7) *for* W, (B1)–(B3) *for* θ, (C1)–(C3) *for* ψ, *and* (D1)–(D3) *for* φ *are satisfied. Suppose that* $\{\lambda^k\}$ *is bounded. If* $p_k \to \infty$ *as* $k \to \infty$ *then the following hold true.*

(i) *Each limit point of the sequence* $\{x^k\}$ *generated by Algorithm* 1 *associated with* L_j $(j = 1, 2, 4)$ *is a global optimal solution to* (P).

(ii) *If the sequence* $\{x_k\}$ *generated by Algorithm* 1 *associated with* L_3 *converges to* $\bar{x}$*, then* $\bar{x}$ *is a global optimal solution to* (P).

We point out that the multiplier updates in Step 2 of Algorithm 1 are not essential for the convergence analysis of Algorithm 1. In fact, no matter how λ^k is updated, Theorem 1 holds true as long as $\{\lambda^k\}$ is bounded. It turns out that this boundedness assumption for $\{\lambda^k\}$ is indispensable for the convergence of Algorithm 1 as shown by the counterexample in [41]. In the subsequent sections, we will discuss several approaches to modify the basic primal-dual scheme by using different strategies. The convergence results will be established for these modified augmented Lagrangian methods without assuming the boundedness of $\{\lambda^k\}$.

4. Modified augmented Lagrangian method using safeguarding. A natural way to relax the boundedness assumption of the Lagrangian multiplier in the basic augmented Lagrangian method is to adopt the safeguarding technique (see [1], [2], [7]) which projects the updated Lagrangian multipliers on suitable bounded intervals.

ALGORITHM 2 (modified primal-dual method using safeguarding).

Step 0 *(initialization). Choose a positive sequence* $\{\epsilon_k\}_{k=1}^{\infty}$ *satisfying* $\epsilon_k \to 0$ *as* $k \to \infty$. *Choose* $\gamma > 1$, $\tau \in (0,1)$, $p_1 > 0$, $\lambda^{\max}$, *and* $\bar{\lambda}^1 \in \mathbb{R}^m$ *such that* $0 < \bar{\lambda}_i^1 \leq \lambda_i^{\max}$ *for* $i = 1, \ldots, m$. *Let* $\sigma^0 = (\sigma_1^0, \ldots, \sigma_m^0)^T$, *with* $\sigma_i^0 = \max\{0, g_i(x^0)\}$, $i = 1, \ldots, m$. *Set* $k = 1$.

Step 1 *(relaxation problem). Compute an* $x^k \in X$ *such that*

$$L_j(x^k, \bar{\lambda}^k, p_k) \leq \min_{x \in X} L_j(x, \bar{\lambda}^k, p_k) + \epsilon_k \qquad (j = 1, 2, 3, 4). \tag{4.1}$$

Step 2 *(multiplier updating). Compute*

$$\lambda_i^{k+1} = \bar{\lambda}_i^k + p_k h_i(x^k, \bar{\lambda}^k, p_k), \qquad i = 1, \ldots, m, \tag{4.2}$$

where h_i *is defined in* (3.1).

Step 3 *(safeguarding projection). Compute*

$$\bar{\lambda}^{k+1} = \mathrm{Proj}_T(\lambda^{k+1}), \tag{4.3}$$

where $\mathrm{Proj}_T(\lambda^{k+1})$ *denotes the Euclidean projection of* λ^{k+1} *on* $T = \{\lambda \in \mathbb{R}^m \mid 0 \leq \lambda_i \leq \lambda_i^{\max},\ i = 1, \ldots, m\}$.

Step 4 *(parameter updating). Let* $\sigma^k = h(x^k, \bar{\lambda}^k, p_k)$. *If*

$$\|\sigma^k\| \leq \tau \|\sigma^{k-1}\|, \tag{4.4}$$

set $p_{k+1} = p_k$; *otherwise, set* $p_{k+1} = \gamma p_k$. *Set* $k := k + 1$, *and go to Step* 1.

Remark 2. The projection operation in (4.3) is a safeguarding strategy to ensure the boundedness of the multipliers used in (4.1). As will be shown later in this section, this safeguarding technique guarantees the convergence of the algorithm without appealing to the restrictive assumption that the multiplier sequence generated by the algorithm is bounded. In Step 4, the Euclidean norm of $h(x, \lambda, p)$ is used to measure the progress of feasibility/complementarity of the inequality constraints $g_i(x) \leq 0$, $i = 1, \ldots, m$.

We first discuss the convergence of Algorithm 2 using L_j $(j = 1, 2)$. We have the following result.

THEOREM 2. *Assume that* (A1)–(A7) *for* W *in* L_1 *and* (B1)–(B3) *for* θ *in* L_2 *are satisfied. Let* $\{x^k\}$ *be the sequence generated by Algorithm* 2 *when using* L_j $(j = 1, 2)$. *Then each limit point of* $\{x^k\}$ *is a global optimal solution to problem* (P).

In the following, we will prove only the theorem for the augmented Lagrangian L_1. The proof for L_2 can be constructed by using similar arguments and the properties (B1)–(B3).

We need the following proposition.

PROPOSITION 1 (Proposition 5.1, [6]). *Let* $s \in \mathbb{R}$ *and* $t \geq 0$. *Then*

(i) $W_t'(s,t) \geq s$;

(ii) $sW_s'(s,t) \geq W(s,t) \geq tW_t'(s,t) \geq st$;

(iii) $W_s'(s,t) = t \ \Rightarrow\ s \leq 0,\ st = 0$.

Proof of Theorem 2. Let $\bar{x}$ be a limit point of $\{x^k\}$, and let $\mathcal{K} \subset \{1, 2, \ldots\}$ be such that $\{x^k\} \to \bar{x}$, as $k \to \infty$ and $k \in \mathcal{K}$. By the closedness of X, $\bar{x} \in X$. We consider the following two cases.

Case (i): $p_k \to \infty$ as $k \to \infty$. By Step 3 of the algorithm, $0 \leq \bar{\lambda}^k \leq \lambda^{\max}$ for all k. Using condition (A6) and the fact that $\inf_{s \in \mathbb{R}} W(s,t) \leq 0$ for all $t \geq 0$, we have

$$0 \geq \inf_{s \in \mathbb{R}} W(s, \bar{\lambda}_i^k) > -\infty \ \text{ for all } k \text{ and } i, \tag{4.5}$$

which in turn implies

$$\lim_{k\to\infty} \frac{1}{p_k} \sum_{i=1}^{m} \inf_{s\in\mathbb{R}} W(s, \bar{\lambda}_i^k) = 0. \tag{4.6}$$

By (4.1), we have

$$L_1(x^k, \bar{\lambda}^k, p_k) \le L_1(x, \bar{\lambda}^k, p_k) + \epsilon_k \quad \forall x \in X.$$

Let x^* be a global solution to problem (P). Then

$$\begin{aligned} L_1(x^k, \bar{\lambda}^k, p_k) &\le L_1(x^*, \bar{\lambda}^k, p_k) + \epsilon_k \\ &= f(x^*) + \frac{1}{p_k} \sum_{i=1}^{m} W(p_k g_i(x^*), \bar{\lambda}_i^k) + \epsilon_k. \end{aligned} \tag{4.7}$$

Note from (2.2) that $W(s,t)$ is an increasing function of s on $\mathbb{R}$ for any fixed $t \ge 0$. Since $g_i(x^*) \le 0$ and $W(0,t) = 0$ for $t \ge 0$, we have

$$W(p_k g_i(x^*), \lambda_i^k) \le W(0, \bar{\lambda}_i^k) = 0$$

for all i and k. Therefore, from (4.7) and the definition of L_1, we have

$$f(x^k) + \frac{1}{p_k} \sum_{i=1}^{m} W(p_k g_i(x^k), \bar{\lambda}_i^k) \le f(x^*) + \epsilon_k. \tag{4.8}$$

Now we are going to show that $\bar{x}$ is feasible to problem (P), i.e., $g_i(\bar{x}) \le 0$, $i = 1, \ldots, m$. Suppose on the contrary that there exists some i_0 such that $g_{i_0}(\bar{x}) > 0$. Let $\epsilon = g_{i_0}(\bar{x})/2$. By the continuity of g_{i_0}, there exists $k_0 > 0$ such that $g_{i_0}(x^k) \ge \epsilon$ for all $k \ge k_0$, $k \in \mathcal{K}$. Since $W(\cdot, t)$ is increasing on $\mathbb{R}$ for any fixed $t \ge 0$ and $\bar{\lambda}^k \ge 0$ for all k, we get

$$W(p_k g_{i_0}(x^k), \bar{\lambda}_{i_0}^k) \ge W(p_k \epsilon, \bar{\lambda}_{i_0}^k) \ge W(p_k \epsilon, 0), \qquad k \ge k_0,\ k \in \mathcal{K}, \tag{4.9}$$

where the second inequality follows from the fact that $W(s, \cdot)$ is an increasing function of $t \ge 0$ for any fixed $s \ge 0$ (cf. Proposition 1(i)). Using (4.9), $\epsilon_k \to 0$, $p_k \to \infty$, and Assumption 1, we deduce from (4.8) that

$$\begin{aligned} f(x^*) + \epsilon_k &\ge f(x^k) + \frac{1}{p_k} \sum_{i=1}^{m} W(p_k g_i(x^k), \bar{\lambda}_i^k) \\ &= f(x^k) + \frac{1}{p_k} W(p_k g_{i_0}(x^k), \bar{\lambda}_{i_0}^k) + \frac{1}{p_k} \sum_{i \ne i_0} W(p_k g_i(x^k), \bar{\lambda}_i^k) \\ &\ge \underline{f} + \frac{1}{p_k} W(p_k \epsilon, 0) + \frac{1}{p_k} \sum_{i \ne i_0} \inf_{s\in\mathbb{R}} W(s, \bar{\lambda}_i^k) \\ &\to \infty \quad (k \to \infty,\ k \in \mathcal{K}), \end{aligned}$$

where the second term in the last inequality tends to ∞ due to the condition (A7), and by (4.6) the third term in the last inequality tends to 0. This contradiction implies that $g_i(\bar{x}) \le 0$ for $i = 1, \ldots, m$.

Next, we show that $\bar{x}$ is a global optimal solution to problem (P). Taking the limit superior in (4.8) and using $\epsilon_k \to 0$ give

$$f(\bar{x}) + \limsup_{k\to\infty, k\in\mathcal{K}} \frac{1}{p_k} \sum_{i=1}^{m} W(p_k g_i(x^k), \bar{\lambda}_i^k) \le f(x^*). \tag{4.10}$$

Since $\bar{x}$ is feasible, we have $f(x^*) \le f(\bar{x})$. Thus, we obtain from (4.10) that

$$\limsup_{k\to\infty, k\in\mathcal{K}} \frac{1}{p_k} \sum_{i=1}^{m} W(p_k g_i(x^k), \bar{\lambda}_i^k) \le 0. \tag{4.11}$$

On the other hand, we have

$$\frac{1}{p_k} \sum_{i=1}^{m} W(p_k g_i(x^k), \bar{\lambda}_i^k) \ge \frac{1}{p_k} \sum_{i=1}^{m} \inf_{s\in\mathbb{R}} W(s, \bar{\lambda}_i^k) \quad \forall k. \tag{4.12}$$

Taking the limit inferior in (4.12) and using (4.6) give rise to

$$\liminf_{k\to\infty, k\in\mathcal{K}} \frac{1}{p_k} \sum_{i=1}^{m} W(p_k g_i(x^k), \bar{\lambda}_i^k) \ge 0. \tag{4.13}$$

Combining (4.11) and (4.13) yields

$$\lim_{k\to\infty, k\in\mathcal{K}} \frac{1}{p_k} \sum_{i=1}^{m} W(p_k g_i(x^k), \bar{\lambda}_i^k) = 0.$$

Thus, we obtain from (4.10) that

$$f(\bar{x}) \le f(x^*).$$

Therefore, $f(\bar{x}) = f(x^*)$, and $\bar{x}$ is a global minimum of problem (P).

Case (ii): $\{p_k\}$ is bounded as $k \to \infty$. In this case, (4.4) in Step 4 is satisfied at each iteration for sufficiently large k. This, together with $\tau \in (0,1)$, implies that $\|\sigma_k\| \to 0$ $(k \to \infty)$ and that there exists $k_1 > 0$ such that $p_k = p_{k_1}$ for all $k \ge k_1$. Since $\sigma_k = h(x^k, \bar{\lambda}^k, p_k)$, it then follows from the definition of h (cf. (3.1)) that

$$\lim_{k\to\infty} \frac{1}{p_k} \left[W_s'\left(p_k g_i(x^k), \bar{\lambda}_i^k\right) - \bar{\lambda}_i^k \right] = 0, \ i = 1, \ldots, m. \tag{4.14}$$

Since $\{\bar{\lambda}^k\} \subset T$ is bounded, we can assume, without loss of generality, that $\bar{\lambda}^k \to \bar{\lambda} \ge 0$ $(k \to \infty, k \in \mathcal{K})$. From (4.14) and $p_k = p_{k_1}$ for all $k \ge k_1$, we obtain

$$W_s'\left(p_{k_1} g_i(\bar{x}), \bar{\lambda}_i\right) - \bar{\lambda}_i = 0, \ i = 1, \ldots, m, \tag{4.15}$$

which in turn implies by Proposition 1(iii) that

$$g_i(\bar{x}) \le 0, \ \bar{\lambda}_i g_i(\bar{x}) = 0, \ i = 1, \ldots, m. \tag{4.16}$$

On the other hand, by Proposition 1(ii), we get

$$st \le W(s,t) \le sW_s'(s,t) \quad \forall s \in \mathbb{R}, \ t \ge 0.$$

It then follows that, for k large enough,

$$\bar{\lambda}_i^k g_i(x^k) \le \frac{1}{p_k} W(p_k g_i(x^k), \bar{\lambda}_i^k) \le g_i(x^k) W_s'(p_k g_i(x^k), \bar{\lambda}_i^k)$$

for all i. Taking the limit in the above inequality and using (4.15) and (4.16), we obtain

$$\lim_{k\to\infty, k\in\mathcal{K}} \frac{1}{p_k} W(p_k g_i(x^k), \lambda_i^k) = \bar{\lambda}_i g_i(\bar{x}) = 0, \quad i = 1, \ldots, m. \tag{4.17}$$

Using (4.17) and the similar arguments as in Case (i), we can show that $\bar{x}$ is a global optimal solution to problem (P). □

Next, we give the convergence result for Algorithm 2 using augmented Lagrangians L_j $(j = 3, 4)$.

THEOREM 3. *Assume that* (C1)–(C3) *for* ψ *in* L_3 *and* (D1)–(D3) *for* φ *in* L_4 *are satisfied. Let* $\{x^k\}$ *be the sequence generated by Algorithm* 2 *using* L_j $(j = 3, 4)$*. If* $x^k \to \bar{x}$ $(k \to \infty)$*, then* $\bar{x}$ *is a global optimal solution to* (P)*.*

Proof. We prove only the theorem for the case of L_3. The case of L_4 can be proved similarly. Since $\bar{\lambda}^1 > 0$ and $\psi'(t) > 0$ for any t, by (4.2) and (4.3) of Algorithm 2, it holds that $\lambda^k > 0$ and $\bar{\lambda}^k > 0$ for all k. Let x^* be a global optimal solution to (P). From (4.1), we have

$$f(x^k) + \frac{1}{p_k} \sum_{i=1}^m \bar{\lambda}_i^k \psi\left(p_k g_i(x^k)\right) \le f(x^*) + \epsilon_k. \tag{4.18}$$

Suppose that $x^k \to \bar{x}$ $(k \to \infty)$. We consider two cases: (i) $p_k \to \infty$ as $k \to \infty$, and (ii) $\{p_k\}$ is bounded as $k \to \infty$. The proof of the theorem in case (i) is similar to that of Theorem 2. We now consider only case (ii). In this case, (4.4) in Step 4 is always satisfied at each iteration for sufficiently large k. This, together with $\tau \in (0, 1)$, implies that $\|\sigma_k\| = \|h(x^k, \bar{\lambda}^k, p_k)\| \to 0$ $(k \to \infty)$ and $p_k = p_{k_0}$ for all $k \ge k_0$ with some sufficiently large k_0. Hence, it follows from the definition of h (cf. (3.1)) that

$$\lim_{k\to\infty} \bar{\lambda}_i^k \left[\psi'(p_{k_0} g_i(x^k)) - 1\right] = 0, \ i = 1, \ldots, m. \tag{4.19}$$

We now show that $g_i(\bar{x}) \le 0$ for $i = 1, \ldots, m$. Suppose that there exists some i_0 such that $g_{i_0}(\bar{x}) > 0$. Then there exists an integer k_0 such that $g_{i_0}(x^k) > 0$ for all $k \ge k_0$. Since ψ' is a strictly increasing function and $\psi'(0) = 1$, we have $\psi'(p_k g_{i_0}(x^k)) > 1$ for all $k \ge k_0$, and hence by (4.2), we have $\lambda_{i_0}^{k+1} > \bar{\lambda}_{i_0}^k$ for all $k \ge k_0$. Also, from (4.3), it holds that $\bar{\lambda}_{i_0}^k \le \bar{\lambda}_{i_0}^{\max}$ for all k. Thus $\lambda_{i_0}^{\max} \ge \bar{\lambda}_{i_0}^{k+1} > \bar{\lambda}_{i_0}^k > 0$ for all $k \ge k_0$. Therefore, there exists $\bar{\lambda}_{i_0}^* > 0$ such that $\bar{\lambda}_{i_0}^k \to \bar{\lambda}_{i_0}^*$ as $k \to \infty$. It then follows from (4.19) that $\bar{\lambda}_{i_0}^* [\psi'(p_{k_0} g_{i_0}(\bar{x})) - 1] = 0$. Note that $\bar{\lambda}_{i_0}^* > 0$ and $\psi'(0) = 1$. Thus, $g_{i_0}(\bar{x}) = 0$. This gives a contradiction, and hence $g_i(\bar{x}) \le 0$ for $i = 1, \ldots, m$.

Since $\{\bar{\lambda}^k\}$ is bounded, we can suppose that $\bar{\lambda}^k \to \bar{\lambda} \ge 0$ as $k \to \infty$ and $k \in \mathcal{K} \subseteq \{1, 2, \ldots\}$. By (4.19), we have

$$\bar{\lambda}_i \psi'(p_{k_0} g_i(\bar{x})) = \bar{\lambda}_i, \quad i = 1, \ldots, m, \tag{4.20}$$

which in turn implies

$$\bar{\lambda}_i g_i(\bar{x}) = 0, \quad i = 1, \ldots, m. \tag{4.21}$$

Moreover, from the convexity of ψ and the condition $\psi(0) = 0$ and $\psi'(0) = 1$, we have

$$t \le \psi(t) \le \psi'(t) t \quad \forall t \in \mathbb{R}. \tag{4.22}$$

Therefore, we obtain

$$\bar{\lambda}_i^k g_i(x^k) \le \frac{1}{p_k}\bar{\lambda}_i^k \psi\left(p_k g_i(x^k)\right) \le g_i(x^k)\bar{\lambda}_i^k \psi'\left(p_k g_i(x^k)\right)$$

for all i and k. Taking the limit in the above inequality and using (4.20) and (4.21), we obtain

$$\lim_{k\to\infty, k\in\mathcal{K}} \frac{1}{p_k}\sum_{i=1}^m \bar{\lambda}_i^k \psi\left(p_k g_i(x^k)\right) = \sum_{i=1}^m \bar{\lambda}_i g_i(\bar{x}) = 0, \tag{4.23}$$

which, together with (4.18), implies $f(\bar{x}) \le f(x^*)$. Thus, $\bar{x}$ is a global optimal solution of (P). □

From the proof of Theorem 2, we see that the multiplier updates in Step 2 of Algorithm 2 are actually not essential for the convergence of Algorithm 2 when using L_j $(j = 1, 2)$. However, the proof of Theorem 3 does rely on the multiplier update in Step 2 when using L_j $(j = 3, 4)$.

5. Modified augmented Lagrangian method with conditional multiplier updating. In this section, we investigate an alternative strategy to modify the basic augmented Lagrangian algorithm for solving (P). The underlying idea is to modify Step 2 of Algorithm 1 for updating the multipliers: The Lagrangian multipliers are updated only when certain progress for the feasibility/complementarity is achieved.

The following modified augmented Lagrangian method for (P) using L_j $(j = 1, 3, 4)$ is an extension of the algorithm in [9] where the equality-constrained problem is considered.

ALGORITHM 3 (modified primal-dual method with conditional multiplier updating).

Step 0. *Choose positive constants*

$$\mu_0,\ \nu_0,\ \beta_\eta,\ \alpha_\epsilon,\ \beta_\epsilon,\ \tau > 1,\ \gamma_1 \in (0,1),\ \omega \ll 1,\ \alpha_\eta > 0.5.$$

Set the initial penalty parameter $p_0 > 1$, *and let*

$$\alpha_0 = \min\left(\frac{1}{p_0}, \gamma_1\right), \quad \epsilon_0 = \nu_0\left(\alpha_0\right)^{\alpha_\epsilon}, \text{ and } \eta_0 = \mu_0\left(\alpha_0\right)^{\alpha_\eta}.$$

Select an initial multiplier vector $\lambda^0 > 0$. *Set* $k = 0$.

Step 1. *Find an* x^k *satisfying*

$$L_j(x^k, \lambda^k, p_k) \le \min_{x\in X} L_j(x, \lambda^k, p_k) + \epsilon_k \qquad (j = 1, 3, 4). \tag{5.1}$$

If

$$\|h(x^k, \lambda^k, p_k)\| \le \eta_k, \tag{5.2}$$

go to Step 2, *where* h *is defined by* (3.1). *Otherwise, go to Step* 3.

Step 2. *If* $\|h(x^k, \lambda^k, p_k)\| \le \omega$, *stop. Otherwise, set*

$$\begin{cases} \lambda^{k+1} = \lambda^k + p_k h(x^k, \lambda^k, p_k), \\ p_{k+1} = p_k, \\ \alpha_{k+1} = \min(\frac{1}{p_{k+1}}, \gamma_1), \\ \epsilon_{k+1} = \epsilon_k(\alpha_{k+1})^{\beta_\epsilon}, \\ \eta_{k+1} = \eta_k\left(\alpha_{k+1}\right)^{\beta_\eta}. \end{cases} \tag{5.3}$$

Set $k := k + 1$, *and go to Step* 1.

Step 3. *Set*

$$\begin{cases} \lambda^{k+1} = \lambda^k, \\ p_{k+1} = \tau p_k, \\ \alpha_{k+1} = \min(\frac{1}{p_{k+1}}, \gamma_1), \\ \epsilon_{k+1} = \nu_0(\alpha_{k+1})^{\alpha_\epsilon}, \\ \eta_{k+1} = \mu_0\,(\alpha_{k+1})^{\alpha_\eta}. \end{cases} \tag{5.4}$$

Set $k := k + 1$, *and go to Step* 1.

Remark 3. In Step 3 of Algorithm 3, the multipliers are not updated when the progress of feasibility/complementarity of the inequality constraints is not satisfactory. It turns out that this scheme maintains the essential boundedness of the multipliers and thus guarantees the convergence of the algorithm when using L_j $(j = 1, 3, 4)$. Numerical comparison between the augmented Lagrangian methods using safeguarding and the above "updating only necessary" strategy can be found in [1], [7].

In the following convergence analysis for Algorithm 3, we set $\omega = 0$. We have the following lemmas regarding the boundedness of $\{\lambda^k\}$ generated by the above algorithm.

LEMMA 1. *If* $p_k \to \infty$ *when Algorithm* 3 *is executed, then* $\lim_{k\to\infty} \frac{\lambda^k}{\sqrt{p_k}} = 0$.

Proof. The proof is similar to that of Lemma 4.1 in [9], where equality constraints are considered. □

LEMMA 2. *If* $\{p_k\}$ *is bounded when Algorithm* 3 *is executed, then* $\{\lambda^k\}$ *is convergent.*

Proof. If $\{p_k\}$ is bounded, there exists $k_0 > 0$ such that Step 2 is executed at every iteration when $k \geq k_0$. Thus, (5.2) is satisfied when $k \geq k_0$. Moreover, from (5.3), we deduce that $p_k = p_{k_0}$ when $k \geq k_0$ and

$$\lambda_i^k = \lambda_i^{k_0} + p_{k_0} \sum_{l=0}^{k-1} h_i(x^{k_0+l}, \lambda^{k_0+l}, p_{k_0+l}), \quad k \geq k_0, \quad i = 1, \ldots, m. \tag{5.5}$$

From (5.3), we see that $\sum_{k=1}^{\infty} \eta_k$ is convergent. Hence, by (5.2), for each i, $\sum_{k=1}^{\infty} h_i(x^k, \lambda^k, p_k)$ is also convergent. It then follows from (5.5) that $\{\lambda^k\}$ is convergent. □

Based on the above lemmas, we are able to establish the convergence results of Algorithm 3 using L_j $(j = 1, 3, 4)$. We first discuss the Algorithm 3 using L_1. In addition to (A1)–(A7), let $W(s, t)$ in the definition of L_1 satisfy the following condition:

(A8) For any nonnegative sequences $\{t_k\} \subseteq \mathbb{R}$ and positive sequence $\{p_k\} \subseteq \mathbb{R}$, with $p_k \to \infty$ $(k \to \infty)$,

$$\lim_{k\to\infty} \frac{t_k}{\sqrt{p_k}} = 0 \Rightarrow \lim_{k\to\infty} \frac{1}{p_k} \inf_{s\in\mathbb{R}} W(s, t_k) = 0.$$

It can be verified that $W(s, t)$ in many augmented Lagrangian functions classified in the general class L_1 satisfies condition (A8). In particular, for $W(s, t)$ defined in (2.3), suppose that we replace condition (E3) by the following condition:

(E3′) There exists $\alpha > 0$ such that $\phi(s) \geq \alpha s^2/2$ for all $s \in \mathbb{R}$.

Then condition (A8) is satisfied. Clearly (E3′) holds if ϕ is strongly convex. We also note that condition (E3′) implies (E3).

THEOREM 4. *Assume that conditions* (A1)–(A8) *for* $W(s, t)$ *are satisfied. Let* $\{x^k\}$ *be the sequence generated by Algorithm* 3 *associated with* L_1. *Then each limit point of the sequence* $\{x^k\}$ *is a global optimal solution to problem* (P).

Proof. Let x^* be a global solution to problem (P). Similar to the proof of Theorem 2, we have

$$f(x^k) + \frac{1}{p_k}\sum_{i=1}^{m} W(p_k g_i(x^k), \lambda_i^k) \le f(x^*) + \epsilon_k. \tag{5.6}$$

Suppose that $x^k \to \bar{x} \in X$ as $k \to \infty$ and $k \in \mathcal{K} \subseteq \{1, 2, \ldots\}$. We consider the following two cases.

Case (i): $p_k \to \infty$ when the algorithm is executed. By Lemma 1, it holds that $\lim_{k\to\infty} \frac{\lambda^k}{\sqrt{p_k}} = 0$, which, by condition (A8), implies

$$\lim_{k\to\infty} \frac{1}{p_k}\sum_{i=1}^{m} \inf_{s\in\mathbb{R}} W(s, \lambda_i^k) = 0. \tag{5.7}$$

Using similar arguments as in the proof of Theorem 2, we can infer from (5.6) and (5.7) that $\bar{x}$ is a global optimal solution to (P).

Case (ii): $\{p_k\}$ is bounded when the algorithm is executed. In this case, there exists $k_0 \ge 0$ such that Step 2 is executed at each iteration when $k \ge k_0$. By the definition of h for L_1 and (5.3), we have

$$\lambda_i^{k+1} = W_s'(p_k g_i(x^k), \lambda_i^k), \quad k \ge k_0. \tag{5.8}$$

Thus, by (2.2) and $\lambda_i^0 > 0$, we infer that $\lambda^k \ge 0$ for all k, and hence by Lemma 2, $\lambda^k \to \bar{\lambda} \ge 0$. Taking the limit in (5.8), we get

$$W_s'\left(p_{k_0} g_i(\bar{x}), \bar{\lambda}_i\right) = \bar{\lambda}_i, \ i = 1, \ldots, m,$$

which, by Proposition 1(iii), implies

$$g_i(\bar{x}) \le 0, \qquad \bar{\lambda}_i g_i(\bar{x}) = 0, \quad i = 1, \ldots, m. \tag{5.9}$$

Using (5.9), we can show, using a proof similar to that of Theorem 2, that

$$\lim_{k\to\infty, k\in\mathcal{K}} \frac{1}{p_k} W(p_k g_i(x^k), \lambda_i^k) = \bar{\lambda}_i g_i(\bar{x}) = 0, \quad i = 1, \ldots, m. \tag{5.10}$$

Since $\epsilon_k \to 0$, we obtain from (5.6) and (5.10) that $f(\bar{x}) \le f(x^*)$. Thus, $\bar{x}$ is a global optimal solution to (P). □

Next, we discuss the convergence property of Algorithm 3 associated with L_3.

THEOREM 5. *Assume that conditions* (C1)–(C3) *for ψ are satisfied. Let $\{x^k\}$ be the sequence generated by Algorithm* 3 *associated with L_3. If $x^k \to \bar{x}$ $(k \to \infty)$, then $\bar{x}$ is a global optimal solution to problem (P).*

Proof. Let x^* be a global solution to problem (P). From (5.1) in Algorithm 3, we have

$$f(x^k) + \frac{1}{p_k}\sum_{i=1}^{m} \lambda_i^k \psi\left(p_k g_i(x^k)\right) \le f(x^*) + \epsilon_k. \tag{5.11}$$

Suppose that $x^k \to \bar{x} \in X$ $(k \to \infty)$. We consider the following two cases.

Case (i): $p_k \to \infty$ when the algorithm is executed. In this case, Step 3 of the algorithm must be executed for infinite times. Let $\mathcal{K} = \{k_0, k_1, k_2, \ldots\}$ be the set of the indices of the iterations in which Step 3 of the algorithm is executed. Then

$$\lambda^{k_j+1} = \lambda^{k_j} \quad \text{for all } j. \tag{5.12}$$

Let t be such that $k_j < k_j + l < k_{j+1}$ for $1 \le l \le t$. At iteration $k_j + l$, $1 \le l \le t$, Step 2 is executed, and hence, for all j and i,

$$\begin{cases} \lambda_i^{k_{j+1}} = \lambda_i^{k_j+t}\psi'(p_{k_j+t}g_i(x^{k_j+t})), \\ \lambda_i^{k_j+l+1} = \lambda_i^{k_j+l}\psi'(p_{k_j+l}g_i(x^{k_j+l})), \quad 1 \le l < t. \end{cases} \tag{5.13}$$

By Lemma 1, it holds that $\lim_{k\to\infty} \frac{\lambda^k}{\sqrt{p_k}} = 0$, which in turn implies

$$\lim_{k\to\infty} \frac{\lambda^k}{p_k} = 0. \tag{5.14}$$

We prove in the following that $g_i(\bar{x}) \le 0$ for $i = 1, \ldots, m$, which together with $\bar{x} \in X$ implies that $\bar{x}$ is a feasible solution to (P). Suppose, on the contrary, there exists some i_0 such that $g_{i_0}(\bar{x}) \ge \epsilon$ for some $\epsilon > 0$. Then there exists an integer $\bar{k}$ such that $g_{i_0}(x^k) \ge \epsilon/2$ for all $k \ge \bar{k}$. Since ψ is strictly increasing and $\psi(0) = 0$, we have

$$\psi(p_k g_{i_0}(x^k)) \ge \psi(p_k\epsilon/2) > 0, \qquad k \ge \bar{k}. \tag{5.15}$$

Since ψ' is a strictly increasing function and $\psi'(0) = 1$, we have

$$\psi'(p_k g_{i_0}(x^k)) > 1 \ \text{ for all } k \ge \bar{k}. \tag{5.16}$$

Note that $\lambda^k > 0$ for all k. Let j_0 be such that $k_{j_0} \ge \bar{k}$. It then follows from (5.12), (5.13), and (5.16) that $\lambda_{i_0}^{k_{j+1}} > \lambda_{i_0}^{k_j}$ for all $j \ge j_0$. Therefore, for all $j \ge j_0$, $\lambda_{i_0}^{k_j} > \lambda_{i_0}^{k_{j_0}} > 0$. Also, condition (C3) implies that $\underline{\psi} = \min_{t\in\mathbb{R}} \psi(t) > -\infty$. Moreover, by Assumption 1, $\underline{f} = \inf_{x\in X} f(x) > -\infty$. Therefore, for $j \ge j_0$, using (5.14) and (5.15), we obtain from (5.11) that

$$\begin{aligned} f(x^*) + \epsilon_{k_j} &\ge f(x^{k_j}) + (1/p_{k_j}) \sum_{i=1}^m \lambda_i^{k_j}\psi(p_{k_j}g_i(x^{k_j})) \\ &\ge \underline{f} + (1/p_{k_j})\lambda_{i_0}^{k_j}\psi(p_{k_j}\epsilon/2) + \frac{1}{p_{k_j}}\underline{\psi}\sum_{i\ne i_0}\lambda_i^{k_j} \\ &\ge \underline{f} + (1/p_{k_j})\lambda_{i_0}^{k_{j_0}}\psi(p_{k_j}\epsilon/2) + \frac{1}{p_{k_j}}\underline{\psi}\sum_{i\ne i_0}\lambda_i^{k_j} \\ &\to \infty \quad (j \to \infty), \end{aligned}$$

where the second term in the last inequality tends to ∞ since, by condition (C3) and $p_{k_j} \to \infty$, it holds that $\lambda_{i_0}^{k_{j_0}} > 0$ and $(1/p_{k_j})\psi(p_{k_j}\epsilon/2) \to \infty$ $(j \to \infty)$, and the third term in the last inequality tends to 0 due to (5.14). This contradiction indicates that $g_i(\bar{x}) \le 0$ for $i = 1, \ldots, m$.

On the other hand, from (5.11), we have

$$f(x^k) + \frac{1}{p_k}\sum_{i=1}^m \underline{\psi}\lambda_i^k \le f(x^*) + \epsilon_k \quad \forall k.$$

In view of (5.14) and $\epsilon_k \to 0$, taking a limit in the above inequality leads to $f(\bar{x}) \le f(x^*)$. Therefore, $\bar{x}$ is a global optimal solution of problem (P).

Case (ii): $\{p_k\}$ is bounded when the algorithm is executed. In this case, there exists $k_0 \geq 0$ such that Step 2 is executed at each iteration when $k \geq k_0$. This implies that $p_k = p_{k_0}$ and

$$\lambda_i^{k+1} = \lambda_i^k \psi'\left(p_k g_i(x^k)\right), \quad i = 1, \ldots, m, \tag{5.17}$$

when $k \geq k_0$. By Lemma 2, $\{\lambda^k\}$ converges to $\bar{\lambda} \geq 0$. Let $M = \{i \mid \bar{\lambda}_i = 0, i = 1, \ldots, m\}$. We claim that $g_i(\bar{x}) \leq 0$ for any $i \in M$. Suppose that there exists $i_0 \in M$ such that $g_{i_0}(\bar{x}) > 0$. Then there exists an integer $\bar{k}$ such that $g_{i_0}(x^k) > 0$ for all $k \geq \bar{k}$. Since ψ' is increasing and $\psi'(0) = 1$, we from (5.17) have that $\lambda_{i_0}^{k+1} \geq \lambda_{i_0}^k > 0$ for $k \geq \bar{k}$. Hence, $\bar{\lambda}_{i_0} > 0$, contradicting $i_0 \in M$. If $i \notin M$, i.e., $\bar{\lambda}_i > 0$, then we have from (5.17) and $\psi'(0) = 1$ that $g_i(\bar{x}) = 0$. Therefore,

$$g_i(\bar{x}) \leq 0, \ \bar{\lambda}_i g_i(\bar{x}) = 0, \quad i = 1, \ldots, m. \tag{5.18}$$

By the convexity of ψ and condition $\psi(0) = 0$ and $\psi'(0) = 1$, it holds that

$$t \leq \psi(t) \leq \psi'(t)t \quad \forall t \in \mathbb{R}. \tag{5.19}$$

Note also that $\lambda^k > 0$ and $p_k > 1$ for all k. Applying (5.17) and (5.19), we obtain

$$\lambda_i^k g_i(x^k) \leq \frac{1}{p_k}\lambda_i^k \psi\left(p_k g_i(x^k)\right) \leq \lambda_i^k \psi'\left(p_k g_i(x^k)\right) g_i(x^k) = \lambda_i^{k+1} g_i(x^k)$$

for all i and $k \geq k_0$. Taking the limit in the above inequality and using (5.18), we obtain

$$\lim_{k\to\infty} \frac{1}{p_k} \sum_{i=1}^m \lambda_i^k \psi\left(p_k g_i(x^k)\right) = \sum_{i=1}^m \bar{\lambda}_i g_i(\bar{x}) = 0. \tag{5.20}$$

Since $\epsilon_k \to 0$, it follows from (5.11) and (5.20) that $f(\bar{x}) \leq f(x^*)$. Thus, $\bar{x}$ is a global optimal solution to (P). □

Finally, we consider Algorithm 3 associated with L_4. Let condition (D3) for the function φ in (2.7) be replaced by the following condition:

(D3′) $\lim_{t\to\infty} \varphi(-t)/\sqrt{t} = 0$.

We observe that functions $\varphi(t) = -\ln(1-t)$ and $\varphi(t) = 1/(1-t) - 1$ satisfy (D3′). It is clear that condition (D3′) implies (D3).

THEOREM 6. *Assume that conditions* (D1)–(D2) *and* (D3′) *for* φ *are satisfied. Let* $\{x^k\}$ *be the sequence generated by Algorithm* 3 *associated with* L_4. *If* $\{x^k\}$ *converges to* $\bar{x}$, *then* $\bar{x}$ *is a global optimal solution to problem* (P).

Proof. Let x^* be a global optimal solution to (P). By (5.1), we have

$$f(x^k) + \frac{1}{p_k}\sum_{i=1}^m \lambda_i^k \varphi(p_k g_i(x^k)) \leq f(x^*) + \frac{1}{p_k}\sum_{i=1}^m \lambda_i^k \varphi(p_k g_i(x^*)) + \epsilon_k. \tag{5.21}$$

From Step 2 of Algorithm 3 and (3.1), we have

$$\lambda_i^{k+1} = \lambda_i^k \varphi'(p_k g_i(x^k)), \quad i = 1, \ldots, m. \tag{5.22}$$

It is clear that $x^k \in \Omega_{p_k}$ and hence $p_k g_i(x^k) < 1$. By (D1) and (D2), φ is a strictly increasing function on $(-\infty, 1)$. Thus, $\varphi'(p_k g_i(x^k)) > 0$ for any i and k. From Step 0 of Algorithm 3, $\lambda^0 > 0$. Hence, (5.22) implies $\lambda_i^k > 0$ for any i and k. Also, both

$\varphi(0) = 0$ and $g_i(x^*) \le 0$ imply $\varphi(p_k g_i(x^*)) \le 0$. Therefore, we obtain the following from (5.21):

$$f(x^k) + \frac{1}{p_k} \sum_{i=1}^{m} \lambda_i^k \varphi(p_k g_i(x^k)) \le f(x^*) + \epsilon_k. \tag{5.23}$$

Now suppose that $x^k \to \bar{x} \in X$ $(k \to \infty)$. Consider the following two cases.

Case (i): $p_k \to \infty$ when the algorithm is executed. Since $g_i(x^k) < 1/p_k$, we deduce that $\bar{x}$ is a feasible solution of (P). By Lemma 1, it holds that

$$\lim_{k\to\infty} \frac{\lambda^k}{\sqrt{p_k}} = 0. \tag{5.24}$$

Denote $I_1 = \{i \mid g_i(\bar{x}) = 0\}$ and $I_2 = \{i \mid g_i(\bar{x}) < 0\}$. Let $\epsilon > 0$ be a constant. For $i \in I_1$, there exists k_1 such that $g_i(x^k) \ge -\epsilon$ for $k \ge k_1$. Since φ is a strictly increasing function on $(-\infty, 1)$, we have

$$\frac{1}{\sqrt{p_k}} \varphi(p_k g_i(x^k)) \ge \frac{1}{\sqrt{p_k}} \varphi(-\epsilon p_k). \tag{5.25}$$

Taking the limit inferior in (5.25) and using condition (D3′) yield the following:

$$\liminf_{k\to\infty} \frac{1}{\sqrt{p_k}} \varphi(p_k g_i(x^k)) \ge 0. \tag{5.26}$$

For $i \in I_2$, there exists k_2 such that $(3/2)g(\bar{x}) \le g_i(x^k) \le (1/2)g(\bar{x}) < 0$ for $k \ge k_2$. Thus,

$$\frac{1}{\sqrt{p_k}} \varphi((3/2)g(\bar{x})p_k) \le \frac{1}{\sqrt{p_k}} \varphi(p_k g_i(x^k)) \le \frac{1}{\sqrt{p_k}} \varphi((1/2)g(\bar{x})p_k). \tag{5.27}$$

Again, taking the limit in (5.27) and using condition (D3′) yield

$$\lim_{k\to\infty} \frac{1}{\sqrt{p_k}} \varphi(p_k g_i(x^k)) = 0. \tag{5.28}$$

Note that $\lambda_i^k > 0$ for all k. Combining (5.24), (5.26), and (5.28) gives rise to the following:

$$\liminf_{k\to\infty} \frac{1}{p_k} \sum_{i=1}^{m} \lambda_i^k \varphi(p_k g_i(x^k) \ge 0.$$

Since $\epsilon_k \to 0$, the above inequality and (5.23) imply that $f(\bar{x}) \le f(x^*)$, that is, $\bar{x}$ is a global solution to (P).

Case (ii): $\{p_k\}$ is bounded when the algorithm is executed. The proof of the theorem in this case is similar to that of Theorem 5. □

6. Normalized multiplier method. In this section, we discuss another approach of achieving the convergence results of the augmented Lagrangian methods without appealing to the boundedness of the multipliers. The idea behind this modification is to normalize the multipliers in the augmented Lagrangian functions L_1, L_2, and L_4 such that the convergence to a global solution can be guaranteed without the boundedness assumption.

The modified augmented Lagrangian functions L_1, L_2, and L_4 are defined as follows:

$$\bar{L}_1(x,\lambda,p) = f(x) + \frac{1}{p}\sum_{i=1}^{m} W\left(pg_i(x), \frac{\lambda_i}{1+\|\lambda\|}\right), \tag{6.1}$$

$$\bar{L}_2(x,\lambda,p) = f(x) + \frac{1}{p}\sum_{i=1}^{m}\left[\theta\left(pg_i(x) + \frac{\lambda_i}{1+\|\lambda\|}\right)_+ - \theta\left(\frac{\lambda_i}{1+\|\lambda\|}\right)\right], \tag{6.2}$$

$$\bar{L}_4(x,\lambda,p) = \begin{cases} f(x) + \frac{1}{p}\sum_{i=1}^{m}\frac{\lambda_i}{1+\|\lambda\|}\varphi(pg_i(x)), & x \in \Omega_p, \\ \infty, & x \in X \setminus \Omega_p. \end{cases} \tag{6.3}$$

Note that the difference between the modified function $\bar{L}_j$ $(j = 1, 2, 4)$ and the original augmented Lagrangian function L_j $(j = 1, 2, 4)$ is the introduction of the factor $\frac{1}{1+\|\lambda\|}$, which can be viewed as a barrier factor to prevent λ from becoming too large. A similar idea was used in [33] for another type of augmented Lagrangian functions.

We describe now the primal-dual scheme using the modified augmented Lagrangians $\bar{L}_j$ $(j = 1, 2, 4)$.

ALGORITHM 4 (normalized multiplier method). *The algorithm is identical to Algorithm* 1 *except that* L_j $(j = 1, 2, 4)$ *in* (3.2) *of Step* 1 *are replaced by* $\bar{L}_j$ $(j = 1, 2, 4)$ *and Step* 2 *is replaced by the following step:*

Step 2. *The multiplier vector* λ^k *is updated by the following equations* $(i = 1, \ldots, m)$*:*

$$\lambda_i^{k+1} = \begin{cases} W'\left(p_k g_i(x^k), \frac{\lambda_i^k}{1+\|\lambda^k\|}\right) & (\text{for } \bar{L}_1), \\ \lambda_i^k + \left[\theta'\left(p_k g_i(x^k) + \frac{\lambda_i^k}{1+\|\lambda^k\|}\right)_+ - \theta'\left(\frac{\lambda_i^k}{1+\|\lambda^k\|}\right)\right] & (\text{for } \bar{L}_2), \\ \frac{\lambda_i^k}{1+\|\lambda^k\|}\varphi'(p_k g_i(x^k)) & (\text{for } \bar{L}_4). \end{cases} \tag{6.4}$$

The convergence of Algorithm 4 to a global solution of (P) is established in the following theorem.

THEOREM 7. *Assume that* (A1)–(A7) *for* W, (B1)–(B3) *for* θ, *and* (D1)–(D3) *for* φ *are satisfied. Suppose in Algorithm* 4 *that* $p_k \to \infty$ *as* $k \to \infty$. *Then each limit point of the sequence* $\{x^k\}$ *generated by Algorithm* 4 *associated with* $\bar{L}_j$ $(j = 1, 2, 4)$ *is a global optimal solution to* (P).

Proof. We prove only the theorem for $\bar{L}_1$. The cases for $\bar{L}_2$ and $\bar{L}_4$ can be proved similarly. From Step 2, we see that $\lambda^k \geq 0$ for all k. Let $\mu^k = (\mu_1^k, \ldots, \mu_m^k)^T$, with

$$\mu_i^k = \frac{\lambda_i^k}{1+\|\lambda^k\|}, \quad i = 1, \ldots, m.$$

Clearly, $\{\mu^k\}$ is bounded and $\mu^k \geq 0$ for all k. The function $\bar{L}_1$ defined in (6.1) can be rewritten as

$$\bar{L}_1(x,\mu,p) = f(x) + \frac{1}{p}\sum_{i=1}^{m} W\left(pg_i(x), \mu_i\right).$$

By Algorithm 4, we have

$$\bar{L}_1(x^k,\mu^k,p_k) \leq \min_{x\in X} \bar{L}_1(x,\mu^k,p_k) + \epsilon_k. \tag{6.5}$$

Also, since $\{\mu^k\}$ is bounded and $p_k \to \infty$, using property (A6), we have

$$\lim_{k\to\infty} \frac{1}{p_k} \sum_{i=1}^{m} \inf_{s\in\mathbb{R}} W(s, \mu_i^k) = 0. \tag{6.6}$$

Applying (6.5)–(6.6) and similar arguments as in the proof of Theorem 2, we can show that each limit point of the sequence $\{x^k\}$ is a global optimal solution to (P). □

Next, we describe a "local" version of Algorithm 4 in which an approximate stationary point of the augmented Lagrangian relaxation problem is computed at each iteration.

ALGORITHM 5.

Step 0. *Select two positive sequences* $\{p_k\}$ *and* $\{\epsilon_k\}$ *such that* $p_k \to \infty$ *and* $\epsilon_k \to 0$ *as* $k \to \infty$. *Select an initial* $\lambda^0 = (\lambda_1^0, \ldots, \lambda_m^0)^T \geq 0$. *Set* $k = 0$.

Step 1. *Compute an* $x^k \in X$ *such that*

$$\|\mathcal{P}[x^k - \nabla_x \bar{L}_j(x^k, \lambda^k, p_k)] - x^k\| \leq \epsilon_k, \quad j = 1, 2, 4, \tag{6.7}$$

where $\mathcal{P}$ *is the Euclidean projection operator onto* X.

Step 2. *Update the multiplier vector* λ_i^k $(i = 1, \ldots, m)$ *by the following formula:*

$$\lambda_i^{k+1} = \begin{cases} W_s'\left(p_k g_i(x^k), \frac{\lambda_i^k}{1+\|\lambda^k\|}\right) & (\text{for } \bar{L}_1), \\ \theta'\left(p_k g_i(x^k) + \frac{\lambda_i^k}{1+\|\lambda^k\|}\right)_+ & (\text{for } \bar{L}_2), \\ \frac{\lambda_i^k}{1+\|\lambda^k\|}\varphi'(p_k g_i(x^k)) & (\text{for } \bar{L}_4). \end{cases} \tag{6.8}$$

Step 3. *Set* $k := k + 1$, *and go to Step* 1.

Remark 4. Note that the multiplier updating formula in (6.8) for $\bar{L}_j$ $(j = 1, 2, 4)$ is motivated by recognizing the following fact:

$$\nabla_x \bar{L}_j(x^k, \lambda^k, p_k) = 0 \Rightarrow \nabla_x L(x^k, \lambda^{k+1}) = 0, \quad j = 1, 2, 4,$$

where $L(x, \lambda) = f(x) + \sum_{i=1}^m \lambda_i g_i(x)$.

DEFINITION 1 (see [7]). *A point* $x^* \in X$ *is said to be degenerate if there exists* $\lambda^* \in \mathbb{R}_+^m$ *such that*

$$\sum_{i\in I(x^*)} \lambda_i^* > 0, \quad \mathcal{P}\left[x^* - \sum_{i\in I(x^*)} \lambda_i^* \nabla g_i(x^*)\right] - x^* = 0,$$

where $I(x^*) = \{i : g_i(x^*) \geq 0,\ i = 1, \ldots, m\}$.

Similar to Theorem 2 in [7], we have the following global convergence results for Algorithm 5.

THEOREM 8. *Assume that condition* (A1)–(A7) *for* W, *conditions* (B1)–(B3) *for* θ, *and conditions* (D1)–(D3) *for* φ *are satisfied. Let* x^* *be a limit point of the sequence* $\{x^k\}$ *generated by the Algorithm* 5 *associated with* $\bar{L}_j$ $(j = 1, 2, 4)$. *Then either* x^* *is degenerate or* x^* *is a KKT point of* (P).

Proof. We prove only the theorem for the case of $\bar{L}_1$. The cases for $\bar{L}_2$ and $\bar{L}_4$ can proved using similar arguments. By (6.7) and (6.8), we have

$$\lim_{k\to\infty} \left\| \mathcal{P}\left[x^k - \nabla f(x^k) - \sum_{i=1}^{m} \lambda_i^{k+1} \nabla g_i(x^k)\right] - x^k \right\| = 0. \tag{6.9}$$

Let $\mathcal{K} \subseteq \{1, 2, \ldots\}$ be such that $\{x^k\}_{\mathcal{K}} \to x^* \in X$. We consider the following two cases.

Case (i): $\{\lambda^{k+1}\}_{\mathcal{K}}$ is unbounded. Then there exists an infinite subsequence $\mathcal{K}_1 \subseteq \mathcal{K}$ such that

$$\Lambda^k = \sum_{i=1}^m \lambda_i^{k+1} \to \infty, \quad k \to \infty,\ k \in \mathcal{K}_1.$$

Since $0 \le \frac{\lambda_i^{k+1}}{\Lambda^k} \le 1$, we may assume that

$$\frac{\lambda_i^{k+1}}{\Lambda^k} \to \lambda_i^*, \quad k \to \infty,\ k \in \mathcal{K}_1, \tag{6.10}$$

for $i = 1, \ldots, m$.

On the other hand, since $\Lambda^k \to \infty$, we have $0 < \frac{1}{\Lambda^k} < 1$ for sufficiently large $k \in \mathcal{K}_1$. By the property of Euclidean projection, we deduce from (6.9) that

$$\lim_{k\to\infty, k\in\mathcal{K}_1} \left\| \mathcal{P}\left[x^k - \frac{1}{\Lambda_k}\left(\nabla f(x^k) - \sum_{i=1}^m \lambda_i^{k+1} \nabla g_i(x^k)\right)\right] - x^k \right\| = 0. \tag{6.11}$$

Since $x^k \to x^*$ and $\Lambda^k \to \infty$ as $k \to \infty$, $k \in \mathcal{K}_1$, we obtain the following from (6.10) and (6.11):

$$\mathcal{P}\left[x^* - \sum_{i=1}^m \lambda_i^* \nabla g_i(x^*)\right] - x^* = 0. \tag{6.12}$$

For $i \notin I(x^*)$, $g_i(x^*) < 0$. Since $p_k \to \infty$, we have $p_k g_i(x^k) \to -\infty$ as $k \to \infty$, $k \in \mathcal{K}$. Note that $\{\frac{\lambda_i^k}{1+\|\lambda^k\|}\}$ is bounded. Using condition (A5) and (6.8), we obtain

$$\lim_{k\to\infty, k\in\mathcal{K}} \lambda_i^{k+1} = \lim_{k\to\infty, k\in\mathcal{K}} W'\left(p_k g_i(x^k), \frac{\lambda_i^k}{1+\|\lambda^k\|}\right) = 0.$$

Thus, by (6.10), $\lambda_i^* = 0$ for $i \notin I(x^*)$. Therefore, we obtain from (6.12) that

$$\mathcal{P}\left[x^* - \sum_{i\in I(x^*)} \lambda_i^* \nabla g_i(x^*)\right] - x^* = 0.$$

So, x^* is degenerate.

Case (ii): $\{\lambda^{k+1}\}_{\mathcal{K}}$ is bounded. In this case, there exists an infinite subsequence $\mathcal{K}_2 \subseteq \mathcal{K}$ such that $\lim_{k\to\infty, k\in\mathcal{K}_2} \lambda^{k+1} = \lambda^* \ge 0$. So, taking the limit in (6.9) gives rise to

$$\mathcal{P}\left[x^* - \nabla f(x^*) - \sum_{i=1}^m \lambda_i^* \nabla g_i(x^*)\right] - x^* = 0. \tag{6.13}$$

We claim that $g_i(x^*) \le 0$ for $i = 1, \ldots, m$. Otherwise, if $g_i(x^*) > 0$ for some i, then $p_k g_i(x^k) \to \infty$, $k \to \infty$, $k \in \mathcal{K}$. Since the convexity of $W(\cdot, t)$ and property (A7) imply that $\lim_{s\to\infty} W_s'(s, t) = \infty$ for any fixed $t \in [0, \infty)$, we have

$$\lambda_i^{k+1} = W_s'\left(p_k g_i(x^k), \frac{\lambda_i^k}{1+\|\lambda^k\|}\right) \to \infty, \quad k \to \infty,\ k \in \mathcal{K},$$

contradicting the boundedness of $\{\lambda_i^{k+1}\}$. Thus, x^* is a feasible solution to (P). If $g_i(x^*) < 0$ for some i, then, as in Case (i), we can show that $\lambda_i^* = 0$. Therefore, we have

$$g_i(x^*) \le 0,\ \lambda_i^* g_i(x^*) = 0, \quad i = 1, \ldots, m.$$

This together with (6.13) implies that x^* is a KKT point of (P) and λ^* is the corresponding optimal multiplier vector. □

7. Concluding remarks. Convergence analysis of the augmented Lagrangian methods has not been well established for constrained global optimization. We have advanced the state of the art of this subject by presenting in this paper new convergence results for the augmented Lagrangian methods in the context of constrained global optimization. The convergence to a global optimal solution has been established for the basic primal-dual scheme under standard conditions. One key assumption in the traditional convergence analysis is that the multiplier sequence generated by the algorithm is bounded. This assumption has been indispensable in the previous convergence analysis for many augmented Lagrangian methods. We have showed that this restrictive condition can be circumvented by using a safeguarding strategy. We have further proposed the modified primal-dual scheme using a conditional multiplier updating strategy and the normalized multiplier method which enable us to achieve the convergence results without the boundedness assumption for the multiplier sequence. Finally, we have established the convergence of the normalized multiplier method for a KKT point.

It turns out that there exist difficulties in the convergence analysis for Algorithm 3 under Mangasarian's augmented Lagrangian L_2 and for Algorithm 4 under the normalized exponential-type augmented Lagrangian L_3. It will be interesting to further investigate these problems in our future research. Another interesting research topic is to identify certain classes of nonconvex optimization problems whose augmented Lagrangian relaxation problems can be *globally* solved efficiently.

Acknowledgments. The authors are grateful to the two anonymous referees for their valuable comments and constructive suggestions for improving the paper. In particular, the investigation of the safeguarding strategy for the augmented Lagrangian methods was suggested by one of the anonymous referees.

REFERENCES

[1] R. Andreani, E. G. Birgin, J. M. Martínez, and M. L. Schuverdt, *On Augmented Lagrangian Methods with General Lower-Level Constraints*, Technical report, 2005. To appear in SIAM J. Optim., available in Optimization Online.

[2] R. Andreani, E. G. Birgin, J. M. Martínez, and M. L. Schuverdt, *Augmented Lagrangian Methods Under the Constant Positive Linear Dependence Constraint Qualification*, Technical report, 2006. To appear in Math. Program., http://www.ime.usp.br/~egbirgin/publications/ abms.pdf.

[3] M. C. Bartholomew-Biggs, *Recursive quadratic programming methods based on the augmented Lagrangian function*, Math. Program. Study, 31 (1987), pp. 21–41.

[4] A. Ben-tal and M. Zibulevsky, *Penalty/barrier multiplier methods for convex programming problems*, SIAM J. Optim., 7 (1997), pp. 347–366.

[5] H. P. Benson, *Deterministic algorithm for constrained concave minimization: A unified critical survey*, Naval Res. Logist., 43 (1996), pp. 765–795.

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

[7] E. G. Birgin, R. A. Castillo, and J. M. Martínez, *Numerical comparison of augmented Lagrangian algorithms for nonconvex problems*, Comput. Optim. Appl., 31 (2005), pp. 31–55.
[8] A. R. Conn, N. I. M. Gould, A. Sartenaer, and Ph. L. Toint, *Convergence properties of an augmented Lagrangian algorithm for optimization with a combination of general equality and linear constraints*, SIAM J. Optim., 6 (1996), pp. 674–703.
[9] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, *A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds*, SIAM J. Numer. Anal., 28 (1991), pp. 545–572.
[10] I. Griva and R. Polyak, *A primal-dual nonlinear rescaling method with dynamic scaling parameter update*, Math. Program., 106 (2006), pp. 237–259.
[11] I. Griva and R. Polyak, *Primal-dual nonlinear rescaling method with dynamic scaling parameter update*, Math. Program., 106 (2006), pp. 237–259.
[12] W. W. Hager, *Dual techniques for constrained optimization*, J. Optim. Theory Appl., 55 (1987), pp. 37–71.
[13] M. R. Hestenes, *Multiplier and gradient methods*, J. Optim. Theory Appl., 4 (1969), pp. 303–320.
[14] R. Horst, P. M. Pardalos, and N. V. Thoai, *Introduction to Global Optimization*, Kluwer Academic Publishers, Dordrecht, 2000.
[15] R. Horst and H. Tuy, *Global Optimization: Deterministic Approaches*, 3rd ed., Springer-Verlag, Berlin, 1996.
[16] X. X. Huang and X. Q. Yang, *A unified augmented Lagrangian approach to duality and exact penalization*, Math. Oper. Res., 28 (2003), pp. 524–532.
[17] X. X. Huang and X. Q. Yang, *Further study on augmented Lagrangian duality theory*, J. Global Optim., 31 (2005), pp. 193–210.
[18] K. C. Kiwiel, *On the twice differentiable cubic augmented Lagrangian*, J. Optim. Theory Appl., 88 (1996), pp. 233–236.
[19] K. C. Kiwiel, *Proximal minimization methods with generalized Bregman functions*, SIAM J. Control Optim., 35 (1997), pp. 1142–1168.
[20] B. W. Kort and D. P. Bertsekas, *A new penalty method for constrained minimization*, in Proceedings of the 1972 IEEE Conference on Decision and Control, New Orleans, 1972, pp. 162–166.
[21] B. W. Kort and D. P. Bertsekas, *Combined primal-dual and penalty methods for convex programming*, SIAM J. Control Optim., 14 (1976), pp. 268–294.
[22] R. M. Lewis and V. Torczon, *A globally convergent augmented Lagrangian pattern search algorithm for optimization with general constraints and simple bounds*, SIAM J. Optim., 12 (2002), pp. 1075–1089.
[23] D. Li, *Zero duality gap for a class of nonconvex optimization problems*, J. Optim. Theory Appl., 85 (1995), pp. 309–324.
[24] D. Li, *Saddle point generation in nonlinear nonconvex optimization*, Nonlinear Anal., 30 (1997), pp. 4339–4344.
[25] D. Li and X. L. Sun, *Local convexification of the Lagrangian function in nonconvex optimization*, J. Optim. Theory Appl., 104 (2000), pp. 109–120.
[26] D. Li and X. L. Sun, *Convexification and existence of saddle point in a pth-power reformulation for nonconvex constrained optimization*, Nonlinear Anal., 47 (2001), pp. 5611–5622.
[27] D. Li and X. L. Sun, *Existence of a saddle point in nonconvex constrained optimization*, J. Global Optim., 21 (2001), pp. 39–50.
[28] O. L. Mangasarian, *Unconstrained Lagrangians in nonlinear programming*, SIAM J. Control Optim., 13 (1975), pp. 772–791.
[29] H. Nakayama, H. Sayama, and Y. Sawaragi, *A generalized Lagrangian function and multiplier method*, J. Optim. Theory Appl., 17 (1975), pp. 211–227.
[30] V. H. Nguyen and J. J. Strodiot, *On the convergence rate of a penalty function method of exponential type*, J. Optim. Theory Appl., 27 (1979), pp. 495–508.
[31] G. Di Pillo and L. Grippo, *An exact penalty function method with global convergence properties for nonlinear programming problems*, Math. Program., 36 (1986), pp. 1–18.
[32] G. Di Pillo and L. Grippo, *Exact penalty functions in constrained optimization*, SIAM J. Control Optim., 27 (1989), pp. 1333–1360.
[33] G. Di Pillo and S. Lucidi, *An augmented Lagrangian function with improved exactness properties*, SIAM J. Optim., 12 (2001), pp. 376–406.
[34] E. Polak and A. L. Tits, *A globally convergent, implementable multiplier method with automatic penalty limitation*, Appl. Math. Optim., 6 (1980), pp. 335–360.
[35] R. Polyak, *Modified barrier functions: Theory and methods*, Math. Program., 54 (1992), pp. 177–222.

[36] R. Polyak, *Log-Sigmoid multipliers method in constrained optimization*, Ann. Oper. Res., 101 (2001), pp. 427–460.
[37] R. Polyak, *Nonlinear rescaling vs. smoothing technique in convex optimization*, Math. Program., 92 (2002), pp. 197–235.
[38] R. Polyak and I. Griva, *Primal-dual nonlinear rescaling method for convex optimization*, J. Optim. Theory Appl., 122 (2004), pp. 111–156.
[39] M. J. D. Powell, *A method for nonlinear constraints in minimization problems*, in Optimization, R. Fletcher, ed., Academic Press, New York, 1969, pp. 283–298.
[40] R. T. Rockafellar, *The multiplier method of Hestenes and Powell applied to convex programming*, J. Optim. Theory Appl., 12 (1973), pp. 555–562.
[41] R. T. Rockafellar, *Augmented Lagrange multiplier functions and duality in nonconvex programming*, SIAM J. Control Optim., 12 (1974), pp. 268–285.
[42] R. T. Rockafellar, *Augmented Lagrangians and applications of the proximal point algorithm in convex programming*, Math. Oper. Res., 1 (1976), pp. 97–116.
[43] J. B. Rosen and P. M. Pardalos, *Constrained Global Optimization: Algorithms and Applications*, Springer-Verlag, Berlin, 1987.
[44] A. M. Rubinov, X. X. Huang, and X. Q. Yang, *The zero duality gap property and lower semicontinuity of the perturbation function*, Math. Oper. Res., 27 (2002), pp. 775–791.
[45] A. M. Rubinov and X. Q. Yang, *Lagrange-type Functions in Constrained Non-Convex Optimization*, Kluwer Academic Publishers, Dordrecht, 2003.
[46] X. L. Sun and D. Li, *Valued-estimation function method for constrained global optimization*, J. Optim. Theory Appl., 102 (1999), pp. 385–409.
[47] X. L. Sun, D. Li, and K. I. M. McKinnon, *On saddle points of augmented Lagrangians for constrained nonconvex optimization*, SIAM J. Optim., 15 (2005), pp. 1128–1146.
[48] X. L. Sun, K. I. M. McKinnon, and D. Li, *A convexification method for a class of global optimization problems with applications to reliability optimization*, J. Global Optim., 21 (2001), pp. 185–199.
[49] P. T. Thatch and H. Tuy, *The relief indicator method for constrained global optimization*, Naval Res. Logist., 37 (1990), pp. 473–497.
[50] P. Tseng and D. P. Bertsekas, *On the convergence of the exponential multiplier method for convex programming*, Math. Program., 60 (1993), pp. 1–19.
[51] H. Tuy, *Monotonic optimization: Problems and solution approaches*, SIAM. J. Optim., 11 (2000), pp. 464–494.
[52] Z. K. Xu, *Local saddle points and convexification for nonconvex optimization problems*, J. Optim. Theory Appl., 94 (1997), pp. 739–746.
[53] H. Yamashita, *A globally convergent constrained quasi-Newton method with an augmented Lagrangian type penalty function*, Math. Program., 23 (1982), pp. 75–86.

SIAM J. OPTIM.
Vol. 18, No. 4, pp. 1231–1249

MATHEMATICAL PROGRAMMING PROBLEMS GOVERNED BY NONLINEAR ELLIPTIC PDES*

M. D. VOISEI†

Abstract. The aim of this article is to find explicit necessary conditions for local optimal pairs of problems governed by divergence-type elliptic PDEs, in terms depending on the nonlinearities involved in the cost functional and state equation. Several examples of optimization problems governed by ODEs and PDEs are presented.

Key words. nonlinear programming, quasi-linear elliptic equations, generalized Jacobian, p-Laplacian, minimal surface equation

AMS subject classifications. 49K20, 35J60, 90C46

DOI. 10.1137/050635730

1. Introduction.

Our main purpose in this paper is to find the optimality conditions satisfied by local optimal pairs (x^*, u^*) of the problem

$$\begin{aligned} \text{(P)} \qquad & \text{Minimize } g(x) + h(u), \text{ on all } (x,u) \in H_0^1(\Omega) \times U, \text{ subject to} \\ \text{(NE)} \qquad & -\operatorname{div} a(\nabla x) + \beta(x) = Bu + f \qquad \text{in } \Omega, \\ & x = 0 \qquad \text{on } \partial\Omega. \end{aligned}$$

Here $\Omega \subset \mathbb{R}^N$ is a bounded C^2-domain, U is a Hilbert space or a reflexive Banach space with a separable dual, $g : L^2(\Omega) \to \mathbb{R}$ is Lipschitz continuous near x^*, $h : U \to \mathbb{R} \cup \{\infty\}$ is proper convex lower semicontinuous, $B \in L(U; L^2(\Omega))$, $f \in L^2(\Omega)$, $a : \mathbb{R}^N \to \mathbb{R}^N$ is strongly monotone and Lipschitz continuous, and $\beta : \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous with a small Lipschitz constant or monotone.

There exists an extensive literature on optimality conditions associated to optimal pairs of optimization problems governed by elliptic PDEs. For state and boundary controlled problems in the semilinear elliptic case we mention [2, 4, 5, 8, 15], while the optimal control of differentiable quasi-linear elliptic PDEs has been studied in [6, 7, 11].

Until now, for quasi-linear elliptic PDEs of the form (NE) with Lipschitz continuous nonlinearities a and β, only selective cases of (P) have been studied; $\beta = 0$ or a linear (see, e.g., [3, 17, 18, 19]). The optimality conditions found in these special cases are not clearly expressed in terms of a and β. The nonlinear Lipschitzian a and β case has not been studied in the literature, and this is the goal of the present note. Moreover, the explicit form of our necessary optimality conditions broadens the area of applications to optimal control problems governed by nonlinear divergence-type elliptic PDEs (see section 3 below).

There is a significant difference between the linear and nonlinear "a" cases in problem (P). Recall that by the Helmholtz decomposition $a(\nabla x) = \nabla p + f_0$, where $x \in H_0^1(\Omega)$, $p \in L^2(\Omega)$, $\operatorname{div} f_0 = 0$ (see, e.g., [16, Lemma 2.5.1, p. 81]). For "a" linear we know that $f_0 = 0$ for every x. For a nonlinear a, the state equation (NE) provides information about p only indirectly.

*Received by the editors July 11, 2005; accepted for publication (in revised form) April 21, 2007; published electronically October 17, 2007.
http://www.siam.org/journals/siopt/18-4/63573.html
†Assistant Professor, Department of Mathematics, Towson University, Towson, MD 21252 (mvoisei@yahoo.com).

There are several advances of this paper over previous results. On one hand, our necessary conditions are derived under less restricting hypotheses, they have an explicit form in terms of the generalized Jacobian of a and generalized gradient of β, and the controlling parameters have better regularity (see Theorem 2.1 below). On the other hand, all necessary conditions previously derived in these settings are particularizations of our conditions of optimality (see [3, 18, 19, 20]).

Our main approach to the maximum principle is based on a family of approximation problems governed by linear state equations. The approximative optimality conditions are found via the closed range theorem. Three primary difficulties must be overcome. First, the existence of approximative optimal solutions requires a special penalization of the cost functional. Second, the process of passing to limit demands a good regularity for a and the controlling parameters. Finally, generalized gradients of integral-type functionals need to be expressed in a clear form. All of these can be fulfilled under the present framework and set of assumptions.

For a general Banach space X we denote by $||\cdot||_X$ the norm in X, by $\langle\cdot,\cdot\rangle_{X\times X^*}$ the duality product between X and X^*, and by $(\cdot,\cdot)_X$ the inner product of X whenever X is a Hilbert space.

Recall that, for a linear operator $A : X \to Y$, $N(A) = \{x \in X;\ Ax = 0\}$ denotes the kernel or null space of A, $R(A) = \{y = Ax;\ x \in D(A)\}$ stands for the range of A, and the adjoint of A is the operator $A^* : Y^* \to X^*$ defined by $x^* = A^*y^*$ iff $\langle x, x^*\rangle_{X\times X^*} = \langle Ax, y^*\rangle_{Y\times Y^*}$ for every $x \in X$.

The indicator of a subset $M \subset X$ is given by $I_M(x) = 0$ for $x \in M$ and $I_M(x) = \infty$ otherwise.

The directional derivative of $f : X \to \mathbb{R}$ at $x \in X$ in the direction of $v \in X$ is given by

$$f^\circ(x; v) = \limsup_{\bar{x}\to x, t\downarrow 0} (1/t)(f(\bar{x} + tv) - f(\bar{x})). \tag{1}$$

Throughout this article "∂" will denote the Clarke generalized gradient or the convex subdifferential. We refer to Clarke [9] for the results about generalized gradients of locally Lipschitz functions and to Zălinescu [21] for the results of convex analysis used in this paper.

The plan of the paper is as follows. Section 2 deals with the abstract optimality conditions for optimal pairs of problem (P). In section 3 several examples of problems governed by ODEs and PDEs and diverse uses of the optimality conditions are presented.

2. A problem governed by a divergence-type elliptic PDE. In what follows we study the optimality conditions for local optimal pairs (x^*, u^*) of problem (P) under the assumptions that $N \geq 1$, Ω is a bounded C^2-domain in $\mathbb{R}^N$, U is a Hilbert space or a reflexive Banach space with a separable dual, $g : L^2(\Omega) \to \mathbb{R}$ is Lipschitz continuous near x^*, $h : U \to \mathbb{R}\cup\{\infty\}$ is proper convex lower semicontinuous, $B : U \to L^2(\Omega)$ is linear bounded, $f \in L^2(\Omega)$, $a : \mathbb{R}^N \to \mathbb{R}^N$ is strongly monotone and Lipschitz continuous; i.e., there exist $L \geq K > 0$ such that for every $r_1, r_2 \in \mathbb{R}^N$

$$(a(r_1) - a(r_2), r_1 - r_2)_{\mathbb{R}^N} \geq K\|r_1 - r_2\|^2_{\mathbb{R}^N}, \tag{2}$$

$$\|a(r_1) - a(r_2)\|_{\mathbb{R}^N} \leq L\|r_1 - r_2\|_{\mathbb{R}^N}, \tag{3}$$

and $\beta : \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous; that is, for some $C_\beta > 0$,

$$|\beta(v_1) - \beta(v_2)| \leq C_\beta|v_1 - v_2|,\ v_1, v_2 \in \mathbb{R}. \tag{4}$$

In addition, we assume that one of the following hypotheses holds:

(H1) $K\lambda_1 > C_\beta$, where $\lambda_1 = \inf\{\frac{\int_\Omega \|\nabla x\|^2_{\mathbb{R}^N} d\omega}{\int_\Omega x^2 d\omega};\ x \in H^1_0(\Omega), x \neq 0\}$ is the first eigenvalue of $-\Delta$ in Ω, with zero Dirichlet boundary condition, or

(H2) β is monotone, i.e., $(\beta(v_1) - \beta(v_2))(v_1 - v_2) \geq 0$, for every $v_1, v_2 \in \mathbb{R}$.

We may assume, without loss of generality, that $\beta(0) = 0$ and $a(0) = 0$. Also, we implicitly understand that the set of admissible pairs is nonempty; i.e., at least (x^*, u^*) satisfies (NE) and $u^* \in \mathrm{dom}\,(h)$ (the domain of h).

As we will see in the subsequent Proposition 2.2, conditions (H1) or (H2) ensure the strong monotonicity of the operator $-\mathrm{div}\ a(\nabla\cdot) + \beta$. This allows us to consider problem (P) as a particular case of the general strongly monotone case studied in [19]. The main improvement provided by the present approach is that, for this specific problem, the optimality conditions can be precisely expressed in terms of the general Jacobian of a and the generalized gradient of β. Some reasons why the condition explicit form is important for applications are that numerical schemes can be built upon the optimality conditions and that the study of qualitative behavior of solutions for certain differential equations is enhanced (see section 3 below).

The analysis of the regularity for the solutions of (NE) as well as the structure of (NE) is needed for future purposes. It is easily seen that the operator $A = -\mathrm{div}\ a(\nabla\cdot) : H^1_0(\Omega) \to H^{-1}(\Omega)$, given by

$$\langle Ax, \varphi\rangle_{H^1_0(\Omega)\times H^{-1}(\Omega)} = \int_\Omega (a(\nabla x), \nabla\varphi)_{\mathbb{R}^N} d\omega,\ \varphi \in H^1_0(\Omega),\ x \in H^1_0(\Omega),$$

is maximal strongly monotone, Lipschitz continuous, invertible, with Lipschitz continuous inverse A^{-1}. The restriction of A to $L^2(\Omega)$

$$A_{L^2} : D(A_{L^2}) = \{x \in H^1_0(\Omega); Ax \in L^2(\Omega)\} \subset L^2(\Omega) \to L^2(\Omega)$$

given by $A_{L^2}x = Ax$, $x \in D(A_{L^2})$, is maximal strongly monotone. The realization $\beta : L^2(\Omega) \to L^2(\Omega)$ is C_β-Lipschitz continuous and monotone whenever β is monotone. Under these terms (NE) has the form

$$A_{L^2}\, x + \beta(x) = Bu + f \text{ in } L^2(\Omega). \tag{5}$$

LEMMA 2.1. *If $a : \mathbb{R}^N \to \mathbb{R}^N$ is strongly monotone and Lipschitz continuous, then*

(a) $D(A_{L^2}) = H^1_0(\Omega) \cap H^2(\Omega)$.

If, in addition, $\nabla^2 a = \left(\dfrac{\partial^2 a}{\partial r_i \partial r_j}\right)_{i,j=1}^N \in L^\infty(\mathbb{R}^N; \mathbb{R}^{N^3})$, *then*

(b) $A : H^1_0(\Omega) \cap H^2(\Omega) \to L^2(\Omega)$ *is locally Lipschitz continuous.*

Proof. (a) Since a is Lipschitz continuous, the superposition

$$a : \mathbb{H}^1(\Omega) := (H^1(\Omega))^N \to \mathbb{H}^1(\Omega),\ a(w) = a \circ w,\ w \in \mathbb{H}^1(\Omega),$$

is well-defined and continuous (see, e.g., [14, Theorem 1, p. 219]).

This makes $-\mathrm{div}\ a(\nabla\cdot) : H^1_0(\Omega) \cap H^2(\Omega) \to L^2(\Omega)$ well-defined; that is, $H^1_0(\Omega) \cap H^2(\Omega) \subset D(A_{L^2})$. For the converse inclusion consider the approximations

$$a_\epsilon(r) = \int_{\mathbb{R}^N} a(r - \epsilon s)J(s)ds = \int_{\mathbb{R}^N} a(s)J_\epsilon(r - s)ds,\ r \in \mathbb{R}^N, \tag{6}$$

where $J \in C_0^\infty(\mathbb{R}^N)$, $\int_{\mathbb{R}^N} J(s)ds = 1$, $J \geq 0$, $J(r) = 0$ for $\|r\|_{\mathbb{R}^N} \geq 1$, $J_\epsilon(r) = \epsilon^{-N} J(r/\epsilon)$, $r \in \mathbb{R}^N$. Recall that a_ϵ is C^1-smooth strongly monotone and Lipschitz continuous with

$$(a_\epsilon(r_1) - a_\epsilon(r_2), r_1 - r_2)_{\mathbb{R}^N} \geq K\|r_1 - r_2\|^2_{\mathbb{R}^N}, \tag{7}$$

$$\|a_\epsilon(r_1) - a_\epsilon(r_2)\|_{\mathbb{R}^N} \leq L\|r_1 - r_2\|_{\mathbb{R}^N}, \tag{8}$$

for every $r_1, r_2 \in \mathbb{R}^N$, and $\|a_\epsilon(r) - a(r)\|_{\mathbb{R}^N} \leq \epsilon L$, for every $r \in \mathbb{R}^N$, (see, e.g., [1, Lemma 2.18, p. 29]).

Whenever $b \in C^1(\mathbb{R}^N)$ the equation $-\text{div } b(\nabla x) = f$, with $f \in L^2(\Omega)$, has a unique solution $x \in H_0^1(\Omega) \cap H^2(\Omega)$ and

$$\|x\|_{H_0^1(\Omega)\cap H^2(\Omega)} \leq C(\|\nabla b\|_{L^\infty})(\|f\|_{L^2(\Omega)} + 1), \tag{9}$$

where $C = C(\|\nabla b\|_{L^\infty}) > 0$ depends proportionally on $\|\nabla b\|_{L^\infty}$ in the sense that the operator $||\nabla a||_{L^\infty} \to C$ is bounded (see, e.g., [12, Remark, p. 498]).

Consider $x \in D(A_{L^2})$. Let $x_\epsilon \in H_0^1(\Omega) \cap H^2(\Omega)$ be the unique solution of $-\text{div } a_\epsilon(\nabla x_\epsilon) = Ax$. We get the following uniform "a priori" estimate:

$$\|x_\epsilon\|_{H_0^1(\Omega)\cap H^2(\Omega)} \leq C(||Ax||_{L^2(\Omega)} + 1),$$

with $C < \infty$, because $\|\nabla a_\epsilon\|_{L^\infty} \leq \|\nabla a\|_{L^\infty} \leq L$ for every $\epsilon > 0$.

This shows that, eventually on a subnet, $x_\epsilon \to x_0$, weakly in $H_0^1(\Omega) \cap H^2(\Omega)$, strongly in $H_0^1(\Omega)$, $\nabla x_\epsilon \to \nabla x_0$, strongly in $L^2(\Omega)$, and $a_\epsilon(\nabla x_\epsilon) \to a(\nabla x_0)$, strongly in $L^2(\Omega)$, by the Lebesgue dominated convergence theorem.

We find that x_0 satisfies $-\text{div } a(\nabla x_0) = Ax$ in $H_0^1(\Omega)$; hence, $x = x_0 \in H_0^1(\Omega) \cap H^2(\Omega)$ and $\|x\|_{H_0^1(\Omega)\cap H^2(\Omega)} \leq C(||Ax||_{L^2(\Omega)}+1)$. Therefore $D(A_{L^2}) \subset H_0^1(\Omega)\cap H^2(\Omega)$.

(b) Condition $\nabla^2 a \in L^\infty(\mathbb{R}^N; \mathbb{R}^{N^3})$ is equivalent to

$$\|\nabla a(r_1) - \nabla a(r_2)\|_{\mathbb{R}^{N^2}} \leq M\|r_1 - r_2\|_{\mathbb{R}^N} \tag{10}$$

for every $r_1, r_2 \in \mathbb{R}^N$, where $M = \|\nabla^2 a\|_{L^\infty} < \infty$.

Since $a_\epsilon \in C^1(\mathbb{R}^N)$ for every $w \in \mathbb{H}^1(\Omega)$, the following chain rule holds:

$$\nabla(a_\epsilon(w)) = \nabla a_\epsilon(w) \cdot \nabla w, \text{ a.e. in } \Omega, \ \epsilon > 0. \tag{11}$$

Using the identity

$$\nabla(a_\epsilon(w_1)) - \nabla(a_\epsilon(w_2)) = [\nabla a_\epsilon(w_1) - \nabla a_\epsilon(w_2)] \cdot \nabla w_1 + \nabla a_\epsilon(w_2) \cdot [\nabla w_1 - \nabla w_2]$$

we find

$$\|\nabla(a_\epsilon(w_1)) - \nabla(a_\epsilon(w_2))\|_{\mathbb{L}^2} \leq \|\nabla^2 a_\epsilon\|_{L^\infty}\|w_1 - w_2\|_{\mathbb{L}^2}\|w_1\|_{\mathbb{H}^1} + L\|w_1 - w_2\|_{\mathbb{H}^1}$$

for every $w_1, w_2 \in \mathbb{R}^N$, where $\mathbb{L}^2(\Omega) := (L^2(\Omega))^N$.

Notice that $\nabla^2 a \in L^\infty(\mathbb{R}^N; \mathbb{R}^{N^3})$ implies $\nabla^2 a_\epsilon \in L^\infty(\mathbb{R}^N; \mathbb{R}^{N^3})$, with

$$\|\nabla^2 a_\epsilon\|_{L^\infty} \leq \|\nabla^2 a\|_{L^\infty} \text{ for every } \epsilon > 0.$$

From $\|\nabla(a(w_1)) - \nabla(a(w_2))\|_{\mathbb{L}^2} = \lim_{\epsilon\to 0} \|\nabla(a_\epsilon(w_1)) - \nabla(a_\epsilon(w_2))\|_{\mathbb{L}^2}$ we get

$$\|\nabla(a(w_1)) - \nabla(a(w_2))\|_{\mathbb{L}^2} \leq \|\nabla^2 a\|_{L^\infty}\|w_1 - w_2\|_{\mathbb{L}^2}\|w_1\|_{\mathbb{H}^1} + L\|w_1 - w_2\|_{\mathbb{H}^1} \tag{12}$$

for every $w_1, w_2 \in \mathbb{R}^N$; i.e., $a : \mathbb{H}^1(\Omega) \to \mathbb{H}^1(\Omega)$ is locally Lipschitz continuous. Taking into account that $-\mathrm{div} : \mathbb{H}^1(\Omega) \to L^2(\Omega)$ is bounded this yields that $A : H_0^1(\Omega) \cap H^2(\Omega) \to L^2(\Omega)$ is locally Lipschitz continuous. □

PROPOSITION 2.1. *Suppose that all of the above assumptions, including* (H1) *or* (H2)*, hold. Then, for every* $f \in L^2(\Omega)$*, the problem*

$$-\mathrm{div}\, a(\nabla x) + \beta(x) = f \text{ in } \Omega, \quad x = 0 \text{ on } \partial\Omega, \tag{13}$$

has a unique solution $x \in H_0^1(\Omega) \cap H^2(\Omega)$.

Proof. From Lemma 2.1 we know that $D(A_{L^2}) \subset H_0^1(\Omega) \cap H^2(\Omega)$. To conclude it is sufficient to prove that $A_{L^2} + \beta$ is surjective.

We claim that conditions (H1) or (H2) imply that $A_{L^2} + \beta$ is maximal strongly monotone and thus surjective. Clearly, this holds if (H2) is fulfilled. The strong monotonicity also assures the uniqueness of (NE) solutions.

Assume that (H1) holds. For every $\varphi_1, \varphi_2 \in D(A_{L^2})$, we have

$$\begin{aligned}((A_{L^2} + \beta)\varphi_1 - (A_{L^2} + \beta)\varphi_2, \varphi_1 - \varphi_2)_{L^2} &= \langle (A + \beta)\varphi_1 - (A + \beta)\varphi_2, \varphi_1 - \varphi_2 \rangle \\ &= \int_\Omega (a(\nabla\varphi_1) - a(\nabla\varphi_2), \nabla(\varphi_1 - \varphi_2))_{\mathbb{R}^N} d\omega + \int_\Omega (\beta(\varphi_1) - \beta(\varphi_2))(\varphi_1 - \varphi_2) d\omega \\ &\geq K\|\nabla(\varphi_1 - \varphi_2)\|_{\mathbb{L}^2}^2 - C_\beta\|\varphi_1 - \varphi_2\|_{L^2}^2 \geq (K\lambda_1 - C_\beta)\|\varphi_1 - \varphi_2\|_{L^2}^2,\end{aligned}$$

where $\langle \cdot, \cdot \rangle$ stands for the dual product in $H_0^1(\Omega) \times H^{-1}(\Omega)$, and we used the Poincaré inequality $\|\varphi\|_{H_0^1} = \|\nabla\varphi\|_{\mathbb{L}^2} \geq \lambda_1^{1/2}\|\varphi\|_{L^2}$, $\varphi \in H_0^1(\Omega)$.

Therefore $A_{L^2} + \beta$ is strongly monotone. It remains to show that $A_{L^2} + \beta$ is maximal monotone provided that $K\lambda_1 > C_\beta$.

Fix ν such that $C_\beta/\lambda_1 < \nu < K$. We write

$$(A_{L^2} + \beta)x = -\mathrm{div}(a - \nu)(\nabla x) + \nu(-\Delta x - (C_\beta/\nu)x) + (C_\beta x + \beta(x)),$$

$x \in D(A_{L^2})$. The operator $\mathbb{R}^N \ni y \longmapsto a(y) - \nu y$ is maximal strongly monotone, since it is continuous defined everywhere and $\nu < K$. This yields that $H_0^1(\Omega) \ni x \longmapsto -\mathrm{div}(a - \nu)(\nabla x) \in H^{-1}(\Omega)$ is maximal strongly monotone. Similarly $H_0^1(\Omega) \ni x \longmapsto -\Delta x - (C_\beta/\nu)x \in H^{-1}(\Omega)$ is maximal strongly monotone and defined everywhere because $C_\beta/\nu < \lambda_1$. Hence, the sum

$$H_0^1(\Omega) \ni x \longmapsto -\mathrm{div}(a - \nu)(\nabla x) + \nu(-\Delta x - (C_\beta/\nu)x) \in H^{-1}(\Omega)$$

is maximal strongly monotone, making its restriction to $L^2(\Omega)$ maximal monotone. Clearly, $L^2(\Omega) \ni x \longmapsto C_\beta x + \beta(x) \in L^2(\Omega)$ is maximal monotone continuous and defined everywhere. Using a perturbation theorem, we get that $A_{L^2} + \beta$ is maximal monotone in $L^2(\Omega)$. □

Remark 2.1. Without (H1), (H2) the state equation (NE) may not have solutions. Take $n = 1$, $a(r) = r$, $\beta(v) = -v$, $r, v \in \mathbb{R}$, $\Omega = (0, \pi)$, $B \equiv 0$, $f \in L^2(0, \pi)$. Then (NE) becomes

$$-x'' - x = f \text{ in } (0, \pi), \quad x(0) = x(\pi) = 0. \tag{14}$$

For this choice of a, β, and Ω we have $K = C_\beta = \lambda_1 = 1$, and (H1), (H2) fail. The general solution of the differential equation in (14) is

$$x(t) = C_1 \cos t + C_2 \sin t + \int_0^t \sin(t - s) f(s) ds, \ t \in (0, \pi), \tag{15}$$

with $C_{1,2}$ constants. The boundary condition reduces to

$$C_1 = 0,\ \int_0^\pi \sin(t)f(t)dt = 0.$$

Hence (14) has the solution $x(t) = C_2 \sin t + \int_0^t \sin(t-s)f(s)ds$ iff $\int_0^\pi \sin(t)f(t)dt = 0$.

The next proposition describes the generalized gradient of some special Lipschitz-type functionals together with some of its robustness properties and "a priori" estimates. The conclusions of this result form the main tool used in the derivation of optimality conditions via an approximation procedure.

PROPOSITION 2.2. *Let X be a Banach spaces, Y be a Hilbert space, and $f : X \to Y$ be a locally Lipschitz continuous function.*

Define $S : X \times Y \to \mathbb{R}$ by

$$S(x,y) = (1/2)||y - f(x)||_Y^2,\ x \in X,\ y \in Y. \tag{16}$$

For $\delta \in Y$, define $f_\delta : X \to \mathbb{R}$ by $f_\delta(x) = -(\delta, f(x))_Y,\ x \in X$, where $(\cdot,\cdot)_Y$ stands for the inner product of Y. Then

(a) *S is locally Lipschitz continuous in $X \times Y$, and f_δ is locally Lipschitz continuous in X for every $\delta \in Y$;*

(b) *$(\alpha, \delta) \in \partial S(x,y)$ iff $\delta = y - f(x)$ and $\alpha \in \partial f_\delta(x)$;*

(c) *$||\alpha||_{X^*} \le C||\delta||_Y$ for every $(\alpha,\delta) \in \partial S(x,y)$, where C is a Lipschitz constant of f near x;*

(d) *if $\alpha_\epsilon \in \partial f_{\delta_\epsilon}(x_\epsilon)$ for every $\epsilon > 0$, and for $\epsilon \downarrow 0$*

$$\begin{aligned} &(1/\epsilon)\alpha_\epsilon \to \alpha,\ \text{weakly in } X^*,\\ &(1/\epsilon)\delta_\epsilon \to \delta,\ \text{strongly in } Y,\\ &x_\epsilon \to x,\ \text{strongly in } X, \end{aligned}$$

then $\alpha \in \partial f_\delta(x)$;

(e) *if in addition $X = Y$ and f is strongly monotone, i.e., for some $K \ge 0$,*

$$(x_1 - x_2, f(x_1) - f(x_2))_Y \ge K||x_1 - x_2||_Y^2 \text{ for every } x_1, x_2 \in Y, \tag{17}$$

then every $\alpha \in \partial f_\delta(x)$ satisfies

$$(\alpha, -\delta)_Y \ge K||\delta||_Y^2 \text{ and } ||\alpha||_Y \ge K||\delta||_Y. \tag{18}$$

Remark 2.2. Note that if the Hilbert space Y is not identified with its dual Y^*, then we can restate (b) as

$$X^* \times Y^* \ni (\alpha,\delta) \in \partial S(x,y) \text{ iff } \delta = J_Y(y - f(x)),\ \alpha \in \partial f_\delta(x), \tag{19}$$

where $J_Y : Y \to Y^*$ is the dual mapping of Y and $f_\delta(x) = -\langle \delta, f(x)\rangle_{Y \times Y^*}$, $x \in X$.

Proof. The proof of (a) is plain. For every $x, v \in X$, $y, w \in Y$ we have

$$\begin{aligned} S^0(x,y;v,w) &= \limsup_{(\overline{x},\overline{y}) \to (x,y), t \downarrow 0} (1/2t)(||\overline{y} + tw - f(\overline{x} + tv)||_Y^2 - ||\overline{y} - f(\overline{x})||_Y^2)\\ &= \limsup_{\overline{x} \to x, t \downarrow 0} (y - f(x), w - (1/t)[f(\overline{x} + tv) - f(\overline{x})])_Y\\ &= (y - f(x), w)_Y + f^0_{(y-f(x))}(x; v). \end{aligned}$$

Therefore $S^0(x,y;v,w) = (y - f(x), w)_Y + f^0_{(y-f(x))}(x;v)$ for every $x, v \in X$, $y, w \in Y$, and (b) follows.

(c) If $X^* \times Y \ni (\alpha, \delta) \in \partial S(x,y)$, then, according to (b), $\delta = y - f(x)$ and $\alpha \in \partial f_\delta(x)$. From $\alpha \in \partial f_\delta(x)$ we find

$$\langle \alpha, v\rangle_{X\times X^*} \le \limsup_{\overline{x}\to x, t\downarrow 0} -(1/t)(\delta, f(\overline{x}+tv) - f(\overline{x}))_Y \le C||\delta||_Y||v||_X \quad \forall v \in X,$$

where C is a Lipschitz constant of f near x.

This final inequality provides us with $||\alpha||_{X^*} \le C||\delta||_Y$.

(d) Clearly $\alpha_\epsilon \in \partial f_{\delta_\epsilon}(x_\epsilon)$ means

$$\langle \alpha_\epsilon, v\rangle_{X\times X^*} \le \limsup_{\widetilde{x}\to x_\epsilon, t\downarrow 0} -(1/t)(\delta_\epsilon, f(\widetilde{x}+tv) - f(\widetilde{x}))_Y$$

for every fixed $v \in X$. Thus, there exist $t_\epsilon \in (0,\epsilon)$, $\widetilde{x}_\epsilon \in X$ such that $\|\widetilde{x}_\epsilon - x_\epsilon\|_X < \epsilon$ and $\langle \alpha_\epsilon, v\rangle_{X\times X^*} \le -(1/t_\epsilon)(\delta_\epsilon, f(\widetilde{x}_\epsilon + t_\epsilon v) - f(\widetilde{x}_\epsilon))_Y + \epsilon$ for every $\epsilon > 0$.

Hence

$$\begin{aligned}\langle (1/\epsilon)\alpha_\epsilon, v\rangle_{X\times X^*} \le{}& -(1/t_\epsilon)(\delta, f(\widetilde{x}_\epsilon + t_\epsilon v) - f(\widetilde{x}_\epsilon))_Y \\ & - (1/t_\epsilon)((1/\epsilon)\delta_\epsilon - \delta, f(\widetilde{x}_\epsilon + t_\epsilon v) - f(\widetilde{x}_\epsilon))_Y + \epsilon. \end{aligned} \tag{20}$$

Because $(1/\epsilon)\delta_\epsilon \to \delta$ strongly in Y and $((1/t_\epsilon)(f(\widetilde{x}_\epsilon + t_\epsilon v) - f(\widetilde{x}_\epsilon))_\epsilon$ is bounded in Y, we have $\lim_{\epsilon\downarrow 0}(1/t_\epsilon)((1/\epsilon)\delta_\epsilon - \delta, f(\widetilde{x}_\epsilon + t_\epsilon v) - f(\widetilde{x}_\epsilon))_Y = 0$.

Let $\epsilon \downarrow 0$ in (20). We get

$$\begin{aligned}\langle \alpha, v\rangle_{X\times X^*} &= \lim_{\epsilon\downarrow 0}\langle (1/\epsilon)\alpha_\epsilon, v\rangle_{X\times X^*} \le \limsup_{\epsilon\downarrow 0} -(1/t_\epsilon)(\delta, f(\widetilde{x}_\epsilon + t_\epsilon v) - f(\widetilde{x}_\epsilon))_Y \\ &\le \limsup_{\widetilde{x}\to x, t\downarrow 0} -(1/t)(\delta, f(\widetilde{x}+tv) - f(\widetilde{x}))_Y = f^0_\delta(x; v)\end{aligned}$$

for every $v \in X$, since $\widetilde{x}_\epsilon \to x$ strongly in X and $t_\epsilon \downarrow 0$ for $\epsilon \downarrow 0$. This shows that $\alpha \in \partial f_\delta(x)$.

(e) Notice that, by (17), for every $\bar{x} \in X$ we have

$$(1/t)\{f_\delta(\bar{x}+t\delta) - f_\delta(\bar{x})\} = -(1/t^2)(t\delta, f(\bar{x}+t\delta) - f(\bar{x}))_Y \le -K\,||\delta||^2_Y.$$

Hence $f^0_\delta(x;\delta) \le -K\,||\delta||^2_Y$. This is sufficient in order to conclude. □

It is known that $-\mathrm{div} : (L^2(\Omega))^N \to H^{-1}(\Omega)$ is linear bounded and $(-\mathrm{div})^* p = \nabla p = (\frac{\partial p}{\partial \omega_i})_{i=1}^N$, $p \in H^1_0(\Omega)$ (see, e.g., [16, p. 80]).

THEOREM 2.1. *Suppose that all the above assumptions, including* (H1) *or* (H2), *hold and* $\nabla^2 a \in L^\infty(\mathbb{R}^N; \mathbb{R}^{N^3})$. *If* $(x^*, u^*) \in H^1_0(\Omega) \times U$ *is a local solution of* (P), *then there exist* $p \in H^1_0(\Omega) \cap H^2(\Omega)$, $\gamma, l \in L^2(\Omega)$, $\eta \in \mathbb{L}^2(\Omega)$ *such that*

(i) $\partial g(x^*) + l + \gamma \ni 0$,
(ii) $B^* p \in \partial h(u^*)$,
(iii) $l = -\mathrm{div}\,\eta, \eta(\omega) \in \partial a(\nabla x^*(\omega))(\nabla p(\omega))$, *a.e.* ω *in* Ω,
(iv) $\gamma(\omega) \in p(\omega)\partial\beta(x^*(\omega))$, *a.e.* ω *in* Ω.

Here ∂a *stands for the generalized Jacobian of* a, *and* $\partial\beta$ *denotes the generalized gradient of* β.

Proof. Let $d = a - \nu i$, where $i(x) = x$, $x \in \mathbb{R}^N$, and $0 < \nu < K$. Notice that d is invertible, Lipschitz continuous with Lipschitz constant $L + \nu$, and maximal strongly monotone with constant $K - \nu > 0$. We denote by $D = -\mathrm{div}\; d(\nabla\cdot)$.

Consider the following approximations of (P):

$$(\mathrm{P}_\epsilon)\quad \begin{array}{l}\text{(locally) minimize } G(x,y,z)+H(u,v),\\ \text{on all } (x,y,z,u,v)\in H_0^1(\Omega)\times\mathbb{L}^2(\Omega)\times L^2(\Omega)\times U\times L^2(\Omega)=:\mathcal{X},\\ \text{subject to } (f,0)=\mathcal{A}(x,y,z,u,v),\end{array}$$

where $\epsilon>0$, $G: H_0^1(\Omega)\times\mathbb{L}^2(\Omega)\times L^2(\Omega)\to\mathbb{R}$ is defined by

$$G(x,y,z)=g(x)+(1/2\epsilon)\|y-d(\nabla x)\|_{\mathbb{L}^2}^2+(1/2\epsilon)\,\|z-\beta(x)\|_{L^2}^2\,, \tag{21}$$

$x\in H_0^1(\Omega)$, $y\in\mathbb{L}^2(\Omega)$, $z\in L^2(\Omega)$, $H: U\times L^2(\Omega)\to\mathbb{R}\cup\{\infty\}$ is given by

$$H(u,v)=h(u)+(1/2)\,\|u-u^*\|_U^2+(1/2)\|v-D(x^*)\|_{L^2}^2\,, \tag{22}$$

$u\in U$, $v\in L^2(\Omega)$, $\mathcal{A}: H_0^1(\Omega)\times\mathbb{L}^2(\Omega)\times L^2(\Omega)\times U\times L^2(\Omega)\to H^{-1}(\Omega)\times H^{-1}(\Omega)$,

$$\mathcal{A}(x,y,z,u,v)=(-\nu\Delta x+z+v-Bu,\ \operatorname{div} y+v),(x,y,z,u,v)\in\mathcal{X}.$$

Here $-\Delta: H_0^1(\Omega)\to H^{-1}(\Omega)$ is the Laplace operator with a zero Dirichlet boundary condition. Notice that the state equation of (P_ϵ) is exactly

$$-\nu\Delta x+z+v=Bu+f,\ v=-\operatorname{div} y. \tag{23}$$

In the statement of (P_ϵ), by local minimization we understand that $(x,u)\in\mathcal{N}$, where $\mathcal{N}$ is a convex closed and bounded neighborhood of (x^*,u^*) in $L^2(\Omega)\times U$ in which (x^*,u^*) is a solution of (P).

Let $(x_n,y_n,z_n,u_n,v_n)_n$ be a minimizing sequence of (P_ϵ); that is,

$$\inf(\mathrm{P}_\epsilon)\le G(x_n,y_n,z_n)+H(u_n,v_n)\le\inf(\mathrm{P}_\epsilon)+1/n, \tag{24}$$

and $(f,0)=\mathcal{A}(x_n,y_n,z_n,u_n,v_n)$ for every $n\ge1$.

From the local minimization sense in which our problem is understood, we know that $(u_n)_n$ is bounded in U and $(x_n)_n$, $(\beta(x_n))_n$ are bounded in $L^2(\Omega)$. Subsequently from (24) we find that $(z_n)_n$, $(v_n)_n$ are bounded in $L^2(\Omega)$. Using the state equation (23), we see that $(\Delta x_n)_n$ is bounded in $L^2(\Omega)$; that is, $(x_n)_n$ is bounded in $H_0^1(\Omega)\cap H^2(\Omega)$. Again from (24) we get the boundedness of $(y_n)_n$ in $\mathbb{L}^2(\Omega)$. Eventually, on a subnet, denoted for simplicity by the same index, we have

$$(x_n,y_n,z_n,u_n,v_n)_n\to(x,y,z,u,y)\text{ weakly in }\mathcal{H},$$

where $\mathcal{H}=H_0^1(\Omega)\cap H^2(\Omega)\times\mathbb{L}^2(\Omega)\times L^2(\Omega)\times U\times L^2(\Omega)$,

$$x_n\to x\text{ strongly in }H_0^1(\Omega);\ \text{that is, }\nabla x_n\to\nabla x\text{ strongly in }\mathbb{L}^2(\Omega).$$

If we let $n\to\infty$ in (24) then we get that $(x,y,z,u,v)=:(x_\epsilon,y_\epsilon,z_\epsilon,u_\epsilon,v_\epsilon)$ is a solution of (P_ϵ), because G is strongly×weakly×weakly lower semicontinuous in $H_0^1(\Omega)\times\mathbb{L}^2(\Omega)\times L^2(\Omega)$ and H is lower semicontinuous in $U\times L^2(\Omega)$.

Since $(x_\epsilon,y_\epsilon,z_\epsilon,u_\epsilon,v_\epsilon)$ is optimal for (P_ϵ) and $(x^*,d(\nabla x^*),\beta(x^*),u^*,D(x^*))$ satisfies its state equation, we have

$$\begin{aligned}G(x_\epsilon,y_\epsilon,z_\epsilon)+H(u_\epsilon,v_\epsilon)&\le G(x^*,d(\nabla x^*),\beta(x^*))+H(u^*,D(x^*))\\ &=g(x^*)+h(u^*)=\inf(\mathrm{P})<\infty\text{ for every }\epsilon>0.\end{aligned} \tag{25}$$

Similarly, inequality (25) and the state equation $-\nu\Delta x_\epsilon + z_\epsilon + v_\epsilon = Bu_\epsilon + f$ show that $(x_\epsilon, y_\epsilon, z_\epsilon, u_\epsilon, v_\epsilon)_\epsilon$ is bounded in $\mathcal{H}$ and for $\epsilon \downarrow 0$

$$\begin{aligned} y_\epsilon - d(\nabla x_\epsilon) &\to 0 \text{ strongly in } \mathbb{L}^2(\Omega), \\ z_\epsilon - \beta(x_\epsilon) &\to 0 \text{ strongly in } L^2(\Omega). \end{aligned}$$

Eventually on a subnet, we have

$$\begin{aligned} &(x_\epsilon, y_\epsilon, z_\epsilon, u_\epsilon, v_\epsilon) \to (x_0, y_0, z_0, u_0, v_0) \text{ weakly in } \mathcal{H}, \\ &x_\epsilon \to x_0 \text{ strongly in } H_0^1(\Omega), \\ &y_\epsilon \to d(\nabla x_0) = y_0 \text{ strongly in } \mathbb{L}^2(\Omega), \\ &z_\epsilon \to \beta(x_0) = z_0 \text{ strongly in } L^2(\Omega), \\ &v_0 = -\text{div } y_0 = D(x_0), \end{aligned}$$

and $-\text{div } a(\nabla x_0) + \beta(x_0) = Bu_0 + f$. Since (x^*, u^*) is optimal for (P), the last equality allows us to continue (25) by

$$G(x_\epsilon, y_\epsilon, z_\epsilon) + H(u_\epsilon, v_\epsilon) \le g(x^*) + h(u^*) \le g(x_0) + h(u_0) \ \forall \epsilon > 0. \tag{26}$$

By passing to lim sup in (26), we find

$$\limsup_{\epsilon \downarrow 0} (1/2)\{\|v_\epsilon - D(x^*)\|_{L^2}^2 + \|u_\epsilon - u^*\|_U^2\} = 0;$$

that is, $v_\epsilon \to D(x^*)$ strongly in $L^2(\Omega)$ and $u_\epsilon \to u^* = u_0$ strongly in U. But $v_\epsilon \to D(x_0)$ weakly in $L^2(\Omega)$, so $Dx_0 = Dx^*$ and $x_0 = x^*$ since D is invertible. From the state equation $-\nu\Delta x_\epsilon + z_\epsilon + v_\epsilon = Bu_\epsilon + f$, we actually get that $x_\epsilon \to x^*$ strongly in $H_0^1(\Omega) \cap H^2(\Omega)$.

We proved that, eventually on a subnet of $\epsilon \downarrow 0$,

$$(x_\epsilon, y_\epsilon, z_\epsilon, u_\epsilon, v_\epsilon) \to (x^*, d(\nabla x^*), \beta(x^*), u^*, D(x^*)) \text{ strongly in } \mathcal{H}.$$

Problem (P_ϵ) has the form

$$\text{(locally) minimize } \tilde{G}(x, y, z, u, v) + \tilde{H}(x, y, z, u, v) + I_M(x, y, z, u, v),$$

where

$$\tilde{G}(x, y, z, u, v) = G(x, y, z), \ \tilde{H}(x, y, z, u, v) = H(u, v), \ (x, y, z, u, v) \in \mathcal{X},$$

and

$$M = \{(x, y, z, u, v) \in \mathcal{X}; (f, 0) = \mathcal{A}(x, y, z, u, v)\}$$

is a closed affine subset of $\mathcal{X}$, since $\mathcal{A} : \mathcal{X} \to H^{-1}(\Omega) \times H^{-1}(\Omega)$ is linear continuous.

Because local minimum points are critical, $(x_\epsilon, y_\epsilon, z_\epsilon, u_\epsilon, v_\epsilon)$ is optimal for (P_ϵ), and $\tilde{G}$ is locally Lipschitz continuous, we find that

$$0 \in \partial\tilde{G}(x_\epsilon, y_\epsilon, z_\epsilon u_\epsilon, v_\epsilon) + \partial(\tilde{H} + I_M)(x_\epsilon, y_\epsilon, z_\epsilon, u_\epsilon, v_\epsilon) \tag{27}$$

The state equation is solvable for every $u \in U$; i.e., the domain of $\tilde{H}$ $\text{dom}\,(\tilde{H}) = H_0^1(\Omega) \times \mathbb{L}^2(\Omega) \times L^2(\Omega) \times \text{dom}\,(h) \times L^2(\Omega)$ satisfies

$$\text{dom}\,(\tilde{H}) - M = \mathcal{X}. \tag{28}$$

More precisely, for every $(x,y,z,u,v)\in\mathcal{X}$ we can write

$$(x,y,z,u,v)=(x,y,z+B(u^*-u)+f,u^*,v)-(0,0,B(u^*-u)+f,u^*-u,0),$$

with $(x,y,z+B(u^*-u)+f,u^*,v)\in\operatorname{dom}\tilde{H}$, $(0,0,B(u^*-u)+f,u^*-u,0)\in M$.

Condition (28) guarantees that $\partial(\tilde{H}+I_M)=\partial\tilde{H}+\partial I_M$, taking into account that $\tilde{H}$, I_M are convex proper lower semicontinuous (see, e.g., [21, Theorem 2.8.7 (vii), p. 126]).

Also, it is easily checked that $\partial I_M(x,y,z,u,v)=N(\mathcal{A})^{\perp}$, $(x,y,z,u,v)\in\mathcal{X}$.

Since $R(\mathcal{A})=H^{-1}(\Omega)\times H^{-1}(\Omega)$, the closed range theorem provides

$$N(\mathcal{A})^{\perp}=R(\mathcal{A}^*),$$

and the optimality condition (27) reduces to

$$0\in\partial\tilde{G}(x_\epsilon,y_\epsilon,z_\epsilon u_\epsilon,v_\epsilon)+\partial\tilde{H}(x_\epsilon,y_\epsilon,z_\epsilon,u_\epsilon,v_\epsilon)+\mathcal{A}^*(p_\epsilon,q_\epsilon) \tag{29}$$

for some $p_\epsilon,q_\epsilon\in H_0^1(\Omega)$.

We have

$$\begin{gathered}\partial\tilde{G}(x_\epsilon,y_\epsilon,z_\epsilon,u_\epsilon,v_\epsilon)=\partial G(x_\epsilon,y_\epsilon,z_\epsilon)\times\{0\}\times\{0\},\\ \partial\tilde{H}(x_\epsilon,y_\epsilon,z_\epsilon,u_\epsilon,v_\epsilon)=\{0\}\times\{0\}\times\{0\}\times\partial H(u_\epsilon,v_\epsilon),\\ \mathcal{A}^*(p_\epsilon,q_\epsilon)=(-\nu\Delta p_\epsilon,-\nabla q_\epsilon,p_\epsilon,-B^*p_\epsilon,p_\epsilon+q_\epsilon).\end{gathered}$$

Therefore (29) becomes

$$\partial G(x_\epsilon,y_\epsilon,z_\epsilon)+(-\nu\Delta p_\epsilon,-\nabla q_\epsilon,p_\epsilon)\ni 0, \tag{30}$$

$$\partial H(u_\epsilon,v_\epsilon)+(-B^*p_\epsilon,p_\epsilon+q_\epsilon)\ni 0. \tag{31}$$

Let $F(x,y)=(1/2)\left\|y-d(\nabla x)\right\|_{\mathbb{L}^2}^2$, $x\in H_0^1(\Omega)$, $y\in\mathbb{L}^2(\Omega)$.

According to Proposition 2.2(b) we have

$$H^{-1}(\Omega)\times\mathbb{L}^2(\Omega)\ni(\alpha,\delta)\in\partial F(x,y)\text{ iff }\delta=y-d(\nabla x),\ \alpha\in\partial d_\delta(x), \tag{32}$$

where $d_\delta(x)=-(\delta,d(\nabla x))_{\mathbb{L}^2}$, $x\in H_0^1(\Omega)$.

We have $d_\delta=f_\delta\circ\nabla$, where $f_\delta(v)=-(\delta,d(v))_{\mathbb{L}^2}$, $v\in\mathbb{L}^2(\Omega)$.

Since $-\operatorname{div}:\mathbb{L}^2(\Omega)\to H^{-1}(\Omega)$ is the adjoint of $\nabla:H_0^1(\Omega)\to\mathbb{L}^2(\Omega)$, by a well-known chain rule (see, e.g., [9, Theorem 2.3.10, p. 45]), we get

$$\partial d_\delta(x)=\{\alpha=-\operatorname{div}\zeta,\ \zeta\in\partial f_\delta(\nabla x)\},\ x\in H_0^1(\Omega).$$

Hence

$$(\alpha,\delta)\in\partial F(x,y)\text{ iff }\alpha=-\operatorname{div}\zeta,\ \delta=y-d(\nabla x),\ \zeta\in\partial f_\delta(\nabla x). \tag{33}$$

Similarly, for $K(x,z)=(1/2)\left\|z-\beta(x)\right\|_{L^2(\Omega)}^2$, $x,z\in L^2(\Omega)$, we find

$$\partial K(x,z)=\{(\gamma,\mu)\in L^2(\Omega)\times L^2(\Omega);\mu=z-\beta(x)),\gamma\in\partial k_\mu(x)\},$$

with $k_\mu(x)=-(\mu,\beta(x))_{L^2}$, $x,z\in L^2(\Omega)$.

Since $H_0^1(\Omega)$ is dense in $L^2(\Omega)$ we know that

$$\partial_{H_0^1}K(x,z)=\partial K(x,z),x\in H_0^1(\Omega),z\in L^2(\Omega)$$

where $\partial_{H_0^1} K$ is the generalized gradient of K seen as a locally Lipschitz function in $H_0^1(\Omega) \times L^2(\Omega)$ (see [9, Corollary, p. 47]).

From its definition we have

$$\begin{aligned} \partial G(x_\epsilon, y_\epsilon, z_\epsilon) = \partial_{H_0^1} g(x_\epsilon) \times \{0\} \times \{0\} + (1/\epsilon)\partial F(x_\epsilon, y_\epsilon) \times \{0\} \\ + (1/\epsilon)\partial_{H_0^1} K(x_\epsilon, z_\epsilon) \times \{0\}, \end{aligned} \tag{34}$$

where $\partial_{H_0^1} g$ is the generalized gradient of g in $H_0^1(\Omega)$.

Again, $\partial_{H_0^1} g(x_\epsilon) = \partial g(x_\epsilon)$ since $x_\epsilon \in H_0^1(\Omega)$, $H_0^1(\Omega)$ is dense in $L^2(\Omega)$, and g is Lipschitz near x^* in $L^2(\Omega)$.

Relation (30) becomes

$$\partial g(x_\epsilon) + (1/\epsilon)\alpha_\epsilon + (1/\epsilon)\gamma_\epsilon - \nu\Delta p_\epsilon \ni 0, \tag{35}$$

$$(1/\epsilon)\delta_\epsilon - \nabla q_\epsilon = 0, \tag{36}$$

$$(1/\epsilon)\mu_\epsilon + p_\epsilon = 0, \tag{37}$$

where $(\alpha_\epsilon, \delta_\epsilon) \in \partial F(x_\epsilon, y_\epsilon)$; that is, $\delta_\epsilon = y_\epsilon - d(\nabla x_\epsilon)$, $\alpha_\epsilon \in \partial d_{\delta_\epsilon}(x_\epsilon)$, or

$$\alpha_\epsilon = -\text{div } \zeta_\epsilon,\ \zeta_\epsilon \in \partial f_{\delta_\epsilon}(\nabla x_\epsilon),\ \delta_\epsilon = y_\epsilon - d(\nabla x_\epsilon), \tag{38}$$

and $(\gamma_\epsilon, \mu_\epsilon) \in \partial K(x_\epsilon, z_\epsilon)$; that is,

$$\gamma_\epsilon \in \partial k_{\mu_\epsilon}(x_\epsilon),\ \mu_\epsilon = z_\epsilon - \beta(x_\epsilon). \tag{39}$$

Relation (31) reduces to

$$B^* p_\epsilon \in \partial h(u_\epsilon) + u_\epsilon - u^*, \tag{40}$$

$$v_\epsilon - D(x^*) + p_\epsilon + q_\epsilon = 0, \tag{41}$$

where, for simplicity, U is considered to be a Hilbert space. If U is a Banach space with a separable dual, then it can be renormed such that its duality mapping is continuous (see, e.g., [10, p. 50]), and the argument below follows similarly.

Since g is Lipschitz near x^* in $L^2(\Omega)$ and $x_\epsilon \to x^*$ strongly in $L^2(\Omega)$ we know that $\cup_{\epsilon>0} \partial g(x_\epsilon)$ is bounded in $L^2(\Omega)$. Relation (35) yields that $((1/\epsilon)\alpha_\epsilon + (1/\epsilon)\gamma_\epsilon - \nu\Delta p_\epsilon)_\epsilon$ is bounded in $L^2(\Omega)$; i.e., for every $\epsilon > 0$

$$\|(1/\epsilon)\alpha_\epsilon + (1/\epsilon)\gamma_\epsilon - \nu\Delta p_\epsilon\|_{L^2} \le C_0 := \sup\left\{ \|w\|_{L^2};\ w \in \bigcup_{\epsilon>0} \partial g(x_\epsilon) \right\} < \infty. \tag{42}$$

Inequality (42) together with (36) and (41) imply

$$\begin{aligned} &\langle (1/\epsilon)\alpha_\epsilon + (1/\epsilon)\gamma_\epsilon - \nu\Delta p_\epsilon, p_\epsilon \rangle_{H_0^1 \times H^{-1}} = ((1/\epsilon)\alpha_\epsilon + (1/\epsilon)\gamma_\epsilon - \nu\Delta p_\epsilon, p_\epsilon)_{L^2} \\ &\quad \le C_0 \|p_\epsilon\|_{L^2} \le C_0 \|q_\epsilon\|_{L^2} + C_0 C_1 \le C_0 \lambda_1^{-1/2} \|q_\epsilon\|_{H_0^1} + C_0 C_1 \\ &\quad = C_0 \lambda_1^{-1/2} \|\nabla q_\epsilon\|_{\mathbb{L}^2} + C_0 C_1 = C_0 \lambda_1^{-1/2} \|(1/\epsilon)\delta_\epsilon\|_{\mathbb{L}^2} + C_0 C_1 \text{ for every } \epsilon > 0, \end{aligned} \tag{43}$$

where $\|v_\epsilon - D(x^*)\|_{L^2} \le C_1 < \infty$ for every $\epsilon > 0$.

From $\zeta_\epsilon \in \partial f_{\delta_\epsilon}(\nabla x_\epsilon)$, we get according to Proposition 2.2(c) and (e) that

$$\|\zeta_\epsilon\|_{\mathbb{L}^2} \le (L + \nu)\|\delta_\epsilon\|_{\mathbb{L}^2}, \tag{44}$$

$$(\zeta_\epsilon, -\delta_\epsilon)_{\mathbb{L}^2} \ge (K - \nu)\|\delta_\epsilon\|_{\mathbb{L}^2}^2. \tag{45}$$

Moreover, if we identify $H^{-1}(\Omega)$ with $H_0^1(\Omega)$, then we have

$$(-\Delta^{-1}\alpha_\epsilon, \delta_\epsilon) \in \partial F(x_\epsilon, y_\epsilon).$$

Since $x_\epsilon \in H_0^1(\Omega) \cap H^2(\Omega)$, F is Lipschitz continuous in $H_0^1(\Omega) \cap H^2(\Omega) \times \mathbb{L}^2(\Omega)$, and $H_0^1(\Omega) \cap H^2(\Omega)$ is dense in $H_0^1(\Omega)$, we know that

$$(-\Delta^{-1}\alpha_\epsilon, \delta_\epsilon) \in \partial F(x_\epsilon, y_\epsilon) \text{ in } H_0^1(\Omega) \cap H^2(\Omega) \times \mathbb{L}^2(\Omega). \tag{46}$$

Therefore, $(\alpha_\epsilon)_\epsilon \subset L^2(\Omega)$, and due to (35) $(p_\epsilon)_\epsilon \subset H_0^1(\Omega) \cap H^2(\Omega)$. Again, from Proposition 2.2(c) and the fact that $d \circ \nabla : H_0^1(\Omega) \cap H^2(\Omega) \to \mathbb{L}^2(\Omega)$ is Lipschitz continuous with Lipschitz constant $(L+\nu)\lambda_1^{-1/2}$, we obtain

$$\| - \Delta^{-1}\alpha_\epsilon\|_{H_0^1 \cap H^2} = \|\alpha_\epsilon\|_{L^2} \leq (L+\nu)\lambda_1^{-1/2}\|\delta_\epsilon\|_{\mathbb{L}^2}. \tag{47}$$

We get from (41), (38), (36), (45), and (47) that for every $\epsilon > 0$

$$\langle (1/\epsilon)\alpha_\epsilon, p_\epsilon\rangle_{H_0^1 \times H^{-1}} = \langle (1/\epsilon)\alpha_\epsilon, -q_\epsilon\rangle_{H_0^1 \times H^{-1}} + ((1/\epsilon)\alpha_\epsilon, D(x^*) - v_\epsilon)_{L^2}$$

$$\geq ((1/\epsilon)\zeta_\epsilon, -\nabla q_\epsilon)_{\mathbb{L}^2} - C_1\|(1/\epsilon)\alpha_\epsilon\|_{L^2} = (1/\epsilon)^2(\zeta_\epsilon, -\delta_\epsilon)_{\mathbb{L}^2} - C_1\|(1/\epsilon)\alpha_\epsilon\|_{L^2}$$

$$\geq (K-\nu)\|(1/\epsilon)\delta_\epsilon\|_{\mathbb{L}^2}^2 - C_1(L+\nu)\lambda_1^{-1/2}\|(1/\epsilon)\delta_\epsilon\|_{\mathbb{L}^2}. \tag{48}$$

Relation $\gamma_\epsilon \in \partial k_{\mu_\epsilon}(x_\epsilon)$ provides us with the estimate

$$\|\gamma_\epsilon\|_{L^2} \leq C_\beta\|\mu_\epsilon\|_{L^2}. \tag{49}$$

We obtain from (37), (49), (41), and (36) that

$$\langle (1/\epsilon)\gamma_\epsilon, p_\epsilon\rangle_{H_0^1 \times H^{-1}} = ((1/\epsilon)\gamma_\epsilon, p_\epsilon)_{L^2} \geq -C_\beta\|p_\epsilon\|_{L^2}^2$$

$$= -C_\beta\{\|q_\epsilon\|_{L^2}^2 + 2(q_\epsilon, v_\epsilon - D(x^*))_{L^2} + \|v_\epsilon - D(x^*)\|_{L^2}^2\}$$

$$\geq -C_\beta\|q_\epsilon\|_{L^2}^2 - 2C_\beta C_1\|q_\epsilon\|_{L^2} - C_\beta C_1^2$$

$$\geq -C_\beta\lambda_1^{-1}\|q_\epsilon\|_{H_0^1}^2 - 2C_\beta C_1\lambda_1^{-1/2}\|q_\epsilon\|_{H_0^1} - C_\beta C_1^2$$

$$= -C_\beta\lambda_1^{-1}\|(1/\epsilon)\delta_\epsilon\|_{\mathbb{L}^2}^2 - 2C_\beta C_1\lambda_1^{-1/2}\|(1/\epsilon)\delta_\epsilon\|_{\mathbb{L}^2} - C_\beta C_1^2. \tag{50}$$

Multiply $(1/\epsilon)\alpha_\epsilon + (1/\epsilon)\gamma_\epsilon - \nu\Delta p_\epsilon$ by p_ϵ in $L^2(\Omega)$ and take (43), (48), and (50) into account to find

$$(K - C_\beta/\lambda_1 - \nu)\|(1/\epsilon)\delta_\epsilon\|_{\mathbb{L}^2}^2 - C_2\lambda_1^{-1/2}\|(1/\epsilon)\delta_\epsilon\|_{\mathbb{L}^2} - C_3 \leq 0, \tag{51}$$

where $C_2 = C_1(L+\nu+2C_\beta) + C_0$, $C_3 = C_\beta C_1^2 + C_0C_1$.

Assume that (H1) is true. We can pick $\nu > 0$ sufficiently small such that $K - C_\beta/\lambda_1 - \nu > 0$. We infer, from (51), that $((1/\epsilon)\delta_\epsilon)_\epsilon$ is bounded in $\mathbb{L}^2(\Omega)$.

Assume that (H2) holds. According to Proposition 2.2(e), from $\gamma_\epsilon \in \partial k_{\mu_\epsilon}(x_\epsilon)$ we find $(\gamma_\epsilon, \mu_\epsilon)_{L^2} \leq 0$. From (37) we get $(\gamma_\epsilon, p_\epsilon)_{L^2} \geq 0$. Again, we multiply $(1/\epsilon)\alpha_\epsilon + (1/\epsilon)\gamma_\epsilon - \nu\Delta p_\epsilon$ by p_ϵ in $L^2(\Omega)$ and take (43) and (48) into account to obtain the boundedness of $((1/\epsilon)\delta_\epsilon)_\epsilon$ in $\mathbb{L}^2(\Omega)$.

Subsequently, we get that $((1/\epsilon)\alpha_\epsilon)_\epsilon$ is bounded in $L^2(\Omega)$ due to (47), $((1/\epsilon)\zeta_\epsilon)_\epsilon$ is bounded in $\mathbb{L}^2(\Omega)$ due to (44), $(q_\epsilon)_\epsilon$ is bounded in $H_0^1(\Omega)$ due to (36), $(p_\epsilon)_\epsilon$ is bounded in $L^2(\Omega)$ due to (41), $((1/\epsilon)\mu_\epsilon)_\epsilon$, $((1/\epsilon)\gamma_\epsilon)_\epsilon$ are bounded in $L^2(\Omega)$ (see (37) and (49)), and from (42) $(p_\epsilon)_\epsilon$ is bounded in $H_0^1(\Omega) \cap H^2(\Omega)$.

Eventually on a subnet, we may assume that

$$\begin{aligned}
&p_\epsilon \to p, \text{ strongly in } H_0^1(\Omega), \text{ weakly in } H_0^1(\Omega) \cap H^2(\Omega),\\
&q_\epsilon \to -p, \text{ strongly in } L^2(\Omega), \text{ weakly in } H_0^1(\Omega),\\
&(1/\epsilon)\delta_\epsilon \to -\nabla p, \text{ weakly in } \mathbb{L}^2(\Omega) \text{ (see (36))}.
\end{aligned}$$

Clearly

$$\begin{aligned}
&(1/\epsilon)\mu_\epsilon \to -p \text{ strongly in } L^2(\Omega) \text{ (see (37))},\\
&(1/\epsilon)\gamma_\epsilon \to \gamma \text{ weakly in } L^2(\Omega),\\
&(1/\epsilon)\zeta_\epsilon \to \zeta \text{ weakly in } \mathbb{L}^2(\Omega),\\
&(1/\epsilon)\alpha_\epsilon \to \alpha = -\text{div } \zeta \text{ weakly in } L^2(\Omega).
\end{aligned}$$

Let $\epsilon \downarrow 0$ in (35) and (40). We find, according to Proposition 2.2(d), that

$$\partial g(x^*) + \alpha + \gamma - \nu\Delta p \ni 0, \ B^*p \in \partial h(u^*) \tag{52}$$

where

$$\gamma \in \partial k_{(-p)}(x^*), \tag{53}$$

and

$$\zeta \in \partial f_{(-\nabla p)}(\nabla x^*), \ \alpha = -\text{div } \zeta \in \partial d_{(-\nabla p)}(x^*). \tag{54}$$

More precisely, we observe that we cannot apply Proposition 2.2(d) for $\zeta_\epsilon \in \partial f_{\delta_\epsilon}(\nabla x_\epsilon)$ to get (54) since $((1/\epsilon)\delta_\epsilon)_\epsilon$ is only weakly convergent in $\mathbb{L}^2(\Omega)$.

Instead, we use $\alpha_\epsilon \in \partial d_{\delta_\epsilon}(x_\epsilon)$ or $-\Delta^{-1}\alpha_\epsilon \in \partial d_{\delta_\epsilon}(x_\epsilon)$ in $H_0^1(\Omega)$. Again, $-\Delta^{-1}\alpha_\epsilon \in \partial d_{\delta_\epsilon}(x_\epsilon)$ in $H_0^1(\Omega) \cap H^2(\Omega)$ by density.

By Lemma 2.1 $D : H_0^1(\Omega) \cap H^2(\Omega) \to L^2(\Omega)$ is well-defined and locally Lipschitz continuous. This allows us to write, according to (36), that

$$\begin{aligned}
d_{\delta_\epsilon}(x) &= -(\delta_\epsilon, d(\nabla x))_{\mathbb{L}^2} = -\langle \epsilon q_\epsilon, D(x)\rangle_{H_0^1 \times H^{-1}}\\
&= -(\epsilon q_\epsilon, D(x))_{L^2} = g_{(\epsilon q_\epsilon)}(x), \ x \in H_0^1(\Omega) \cap H^2(\Omega),
\end{aligned}$$

where for $q \in L^2(\Omega)$, $g_q(x) = -(q, D(x))_{L^2}$, $x \in H_0^1(\Omega) \cap H^2(\Omega)$.

Now we can use Proposition 2.2(d) because $q_\epsilon \to -p$ strongly in $L^2(\Omega)$.

From

$$-\Delta^{-1}\alpha_\epsilon \in \partial d_{\delta_\epsilon}(x_\epsilon) = \partial g_{(\epsilon q_\epsilon)}(x_\epsilon) \text{ in } H_0^1(\Omega),$$

we find

$$-\Delta^{-1}\alpha \in \partial g_{(-p)}(x^*) = \partial d_{(-\nabla p)}(x^*) \text{ in } H_0^1(\Omega) \cap H^2(\Omega).$$

Equivalently, $-\Delta^{-1}\alpha \in \partial d_{(-\nabla p)}(x^*)$ in $H_0^1(\Omega)$, or

$$\alpha \in \partial d_{(-\nabla p)}(x^*) \text{ in } H_0^1(\Omega) \times H^{-1}(\Omega).$$

This final relation can be restated as $\alpha = -\text{div } \zeta$, with $\zeta \in \partial f_{(-\nabla p)}(\nabla x^*)$, and (54) is proved.

Clearly, $\gamma \in \partial k_{(-p)}(x^*)$ means $\gamma \in \partial \Gamma(x^*)$, where

$$\Gamma(x) = \int_\Omega p(\omega)\beta(x(\omega))d\omega,\ x \in L^2(\Omega),$$

and it can be described as $\gamma(\omega) \in p(\omega)\partial\beta(x^*(\omega))$, a.e. ω in Ω (see [9, Theorem 2.7.2, p. 76]).

The first part in (54) is equivalent to $\eta := \zeta + \nu\nabla p \in \partial\Psi_{(-\nabla p)}(\nabla x^*)$, where

$$\Psi_{(-\nabla p)}(v) = (\nabla p, a(v))_{\mathbb{L}^2} = \int_\Omega (a(v(\omega)), \nabla p(\omega))_{\mathbb{R}^N} d\omega,\ v \in \mathbb{L}^2(\Omega),$$

and it can be described as $\eta(\omega) \in \partial a(\nabla x^*(\omega))(\nabla p(\omega))$, a.e. ω in Ω (see [9, Theorem 2.6.6, p. 72, and Theorem 2.7.2, p. 76]). Take $l := \alpha - \nu\Delta p = -\text{div } \eta$. The proof is complete. □

Remark 2.3. Using a similar argument and slightly different computations one can show that Theorem 2.1 holds if instead of (H1) or (H2) we consider assumptions which ensure that $-\text{div } a(\nabla\cdot) + \beta : H_0^1(\Omega) \to H^{-1}(\Omega)$ is strongly monotone.

COROLLARY 2.1. *Assume that all of the assumptions of Theorem* 2.1 *hold, with the relaxed conditions: a is locally Lipschitz and $\nabla^2 a \in L^\infty_{loc}(\mathbb{R}^N; \mathbb{R}^{N^3})$. In addition, suppose that every solution $x \in H_0^1(\Omega) \cap H^2(\Omega)$ of* (NE) *satisfies the "a priori" estimate*

$$\|\nabla x\|_{L^\infty} \le \gamma = \gamma(\|u\|_U), \tag{55}$$

where $\gamma = \gamma(\|u\|_U)$ is a bounded function of $\|u\|_U$.

Then all of the conclusions of Theorem 2.1 *hold.*

Proof. Since a is locally Lipschitz and $\nabla^2 a \in L^\infty_{loc}(\mathbb{R}^N; \mathbb{R}^{N^3})$ we can modify a outside a bounded subset of $\mathbb{R}^N$ without locally changing the set of solutions of (NE) such that the new a is Lipschitz continuous with $\nabla^2 a \in L^\infty(\mathbb{R}^N; \mathbb{R}^{N^3})$, and we can apply Theorem 2.1 to get the conclusions. □

Remark 2.4. For $N = 1$ we have $\nabla x \in H^1(\Omega) \subset L^\infty(\Omega)$, and from the "a priori" estimate contained in the proof of Lemma 2.1(a) we know that $\|\nabla x\|_{H^1} \le C(\|Bu + f\|_{L^2} + 1)$, from which we find (55). More precisely,

$$\|\nabla x\|_{L^\infty} \le \frac{1}{(K\lambda_1 - C_\beta)\sqrt{6}}\|Bu + f\|_{L^2}.$$

Therefore, in the 1-dimensional case, the conclusions of Theorem 2.1 remain true with the relaxed conditions a locally Lipschitz continuous and $\nabla^2 a \in L^\infty_{loc}(\mathbb{R}^N; \mathbb{R}^{N^3})$.

3. Examples.

Example 3.1. Consider the optimization problem

$$\text{minimize } \frac{1}{2}\int_0^\pi |x(t)|^2 dt + \frac{1}{2}\int_0^\pi |u(t)|^2 dt \tag{56}$$

subject to

$$-x_{tt}(1 + e^{-x_t}) = u + f,\ \text{a.e. in } (0,\pi),\ x(0) = x(\pi) = 0, \tag{57}$$

where $u \in L^2(0,\pi)$.

This problem is of the form (P) with $g(x) = \frac{1}{2}\|x\|^2_{L^2}, h(u) = \frac{1}{2}\|u\|^2_{L^2}$, $x, u \in L^2(0,\pi)$, $a(r) = r - e^{-r}$, $r \in \mathbb{R}$, $\beta = 0$, $Bu = u$, $u \in L^2(0,\pi)$, $f \in L^2(0,\pi)$.

Notice that g, h, a, β, B, f satisfy the assumptions of Theorem 2.1 since a is locally Lipschitz continuous with $K = 1$, a'' is locally bounded (see Remark 2.4), and $C_\beta = 0$, $\lambda_1 = 1$.

Therefore, for every solution (x^*, u^*) there is a $p \in H^1_0(0,\pi) \cap H^2(0,\pi)$, $\gamma, l, \eta \in L^2(0,\pi)$ such that

$$x^* + l + \gamma = 0,\ p = u^*,\ l = -\eta_t,\ \eta = (1 + e^{-x^*_t})p_t,\ \gamma = 0; \tag{58}$$

that is,

$$x^* - ((1 + e^{-x^*_t})u^*_t)_t = 0 \text{ a.e. in } (0,\pi). \tag{59}$$

Combined with (57), relation (59) can answer several questions such as: Is it possible to bring the solutions x of (57) into the origin with minimal L^2-effort in u, or, in other words, is there an optimal pair of our problem of the form $(0, u^*)$? The answer is positive iff $f = 0$ since for $x^* = 0$ we get $u^*_{tt} = 0$ in $H^1_0(0,\pi) \cap H^2(0,\pi)$, i.e., $u^* = 0$, which contradicts (57) if $f \neq 0$.

Moreover, (57) and (59) together with the boundary conditions $x^*(0) = x^*(\pi) = u^*(0) = u^*(\pi) = 0$ can be used to determine (x^*, u^*) numerically.

Example 3.2. Consider the problem

$$\text{minimize } \ell^2 \int_0^T \cos^2\theta(t)dt + \frac{1}{2}\int_0^T |u(t)|^2 dt$$

subject to the pendulum problem

$$\theta_{tt} + \frac{g}{\ell}\sin\theta = u + f,\ \text{a.e. in } (0,T),\ \theta(0) = \theta(T) = 0. \tag{60}$$

Here g is the gravitational acceleration, ℓ is the length of the pendulum, $T > 0$ is fixed, and $f \in L^2(0,T)$.

The objective functional in this problem expresses the goal of maximizing the height of the pendulum $H = \ell - \ell\cos\theta$ with minimum L^2-effort in u, under the condition $\theta(0) = \theta(T) = 0$.

This problem is of type (P) with $g(\theta) = \ell^2 \int_0^T \cos^2\theta(t)dt$, $h(u) = \frac{1}{2}\|u\|^2_{L^2}$, $\theta, u \in L^2(0,T)$, $a(r) = r$, $\beta(r) = -\frac{g}{\ell}\sin r$, $r \in \mathbb{R}$, $Bu = -u$, $u \in L^2(0,T)$. We have $K = L = 1$, $C_\beta = \frac{g}{\ell}$, $\lambda_1 = \frac{\pi^2}{T^2}$. Therefore, for $\ell > \frac{gT^2}{\pi^2}$, we may apply Theorem 2.1 to get that for every optimal pair (θ^*, u^*) there exist $p \in H^1_0(0,T) \cap H^2(0,T)$, $\gamma, l, \eta \in L^2(0,T)$ such that

$$\partial g(\theta^*) + l + \gamma \ni 0,\ -p = u^*,\ l = -\eta_t,\ \eta = p_t,\ \gamma = -p\frac{g}{\ell}\cos\theta^*; \tag{61}$$

that is,

$$-\ell^2\sin(2\theta^*) + u^*_{tt} + \frac{g}{\ell}u^*\cos\theta^* = 0. \tag{62}$$

Various conclusions can be drawn from (60) and (62) including the possibility for numerical determination of optimal pairs. For example, if $(\theta^*(t) = \theta^*, u^*(t) = u(t))$ is a stationary optimal pair, then $\theta^* = \pm\pi/2$, $f(t) = -u^* \pm \frac{g}{\ell}$ or $\sin\theta^* = \frac{g}{2\ell^3}u^*$, $f(t) = (\frac{g^2}{2\ell^4} - 1)u^*$.

Example 3.3. Let $N \geq 1$, $C = (c_{ij})_{1 \leq i,j \leq N} \in \mathcal{M}_N(\mathbb{R})$ be a square positive definite matrix, i.e., $(Cr, r)_{\mathbb{R}^N} \geq c\|r\|^2_{\mathbb{R}^N}$, for every $r \in \mathbb{R}^N$, $c > k > 0$, and

$$a(r) = Cr + k(|r_1|, |r_2|, \dots, |r_N|)^T,\ r = (r_1, r_2, \dots, r_N)^T \in \mathbb{R}^N. \tag{63}$$

Here "^{T}" stands for the transpose of a matrix. Notice that a is Lipschitz continuous, strongly monotone with constant $c - k > 0$, nondifferentiable at 0, and $\nabla^2 a \in L^\infty(\mathbb{R}^N; \mathbb{R}^{N^3})$.

Problem (P) becomes in these settings

$$\begin{aligned} &\text{(P)} \quad \text{minimize } g(x) + h(u) \text{ on all } (x,u) \in H_0^1(\Omega) \times U, \text{ subject to} \\ &\text{(NE)} \quad -\sum_{i,j=1}^N c_{ij} \frac{\partial^2 x}{\partial \omega_i \partial \omega_j} - k \sum_{i=1}^N \operatorname{sign}\left(\frac{\partial x}{\partial \omega_i}\right) \frac{\partial^2 x}{\partial \omega_i^2} + \beta(x) = f \text{ in } \Omega, \\ &\quad x = 0 \text{ on } \partial\Omega. \end{aligned}$$

Whenever β satisfies (H1) or (H2), for every optimal pair (x^*, u^*) of (P) we have

$$\partial g(x^*) + l + \gamma \ni 0,\ B^* p \in \partial h(u^*),\ l = -\operatorname{div} \eta, \tag{64}$$

$$\eta \in \left(\sum_{j=1}^N c_{ij} \frac{\partial p}{\partial \omega_j} + k \operatorname{sign}\left(\frac{\partial x^*}{\partial \omega_i}\right) \frac{\partial p}{\partial \omega_i} \right)^T_{1 \leq i \leq N},\ \gamma \in p\partial\beta(x^*) \text{ a.e. in } \Omega, \tag{65}$$

for some $p \in H_0^1(\Omega) \cap H^2(\Omega)$, $\gamma, l, \eta \in \mathbb{L}^2(\Omega)$. Here $\operatorname{sign} \omega = \omega/|\omega|$, $\omega \neq 0$, $\operatorname{sign} 0 = [-1, 1]$.

Example 3.4. Let $N \geq 2$, $p \geq 3$, and $a(r) = r\|r\|^{p-2}$, $r \in \mathbb{R}^N$. In this case (P) has the form

$$\begin{aligned} &\text{(P)} \quad \text{minimize } g(x) + h(u) \text{ on all } (x,u) \in H_0^1(\Omega) \times U, \text{ subject to} \\ &\qquad \text{(NE)} \ -\ \Delta_p x + \beta(x) = Bu + f \ \text{ in } H_0^1(\Omega), \end{aligned}$$

where "Δ_p" denotes the p-Laplacian.

More precisely, $\Delta_p : W_0^{1,p}(\Omega) \to W^{-1,q}(\Omega)$, $-\Delta_p x = -\operatorname{div}(\nabla x \|\nabla x\|^{p-2}_{\mathbb{R}^N}) = g \in W^{-1,q}(\Omega)$ iff $x \in W_0^{1,p}(\Omega)$, and

$$\int_\Omega (\|\nabla x\|^{p-2}_{\mathbb{R}^N} \nabla x \cdot \nabla \varphi - g\varphi) d\omega = 0 \text{ for every } \varphi \in C_0^\infty(\Omega),$$

or, equivalently, x is a minimizer of the functional

$$E : W_0^{1,p}(\Omega) \to \mathbb{R},\ E(v) = \frac{1}{p} \int_\Omega \|\nabla v\|^p_{\mathbb{R}^N} d\omega - \int_\Omega gv\, d\omega,\ v \in W_0^{1,p}(\Omega).$$

The p-Laplacian operator arises in various physical contexts: nonlinear elasticity, reaction-diffusion problems, non-Newtonian fluid flow models, diffusive logistic equation, nonlinear elastic membranes, electrochemical machining, elastic-plastic torsional creep, the Monge mass transfer problem, etc.

We take $f \in L^\infty(\Omega)$ and $B \in L(U; L^\infty(\Omega))$.

A direct computation shows that, for $p \geq 3$, $\nabla^2 a \in L^\infty_{loc}(\mathbb{R}^N; \mathbb{R}^{N^3})$, and since a is locally Lipschitz continuous it is sufficient to show (55) in order to apply Corollary 2.1.

We take β with a sufficiently small Lipschitz constant or monotone, and after we multiply (NE) by x we find

$$\begin{aligned} -\Delta_p x(\omega)x(\omega) &\le -\beta(x(\omega))x(\omega) + |(Bu+f)(\omega)||x(\omega)| \\ &\le C^*_\beta |x(\omega)|^2 + ||Bu+f||_{L^\infty}|x(\omega)| \text{ a.e. } \omega \in \Omega, \end{aligned} \tag{66}$$

where $C^*_\beta = 0$ if β is monotone and $C^*_\beta = C_\beta$ if β is Lipschitz.

After integration over Ω we get

$$\|x\|^{p-1}_{W^{1,p}} \le \text{const}\|Bu+f\|_{L^2}. \tag{67}$$

According to [13, Theorem 1.5.5, p. 115] together with the Sobolev embedding $W^{1,p}_0(\Omega) \subset L^{p_0}(\Omega)$, where $p_0 = Np/(N-p)$ if $p < N$, $p_0 = 2p$ if $p \ge N$, relations (66), (67) imply $\|x\|_{L^\infty} \le \gamma(\|u\|_U)$, which, according to [13, Theorem 1.5.6, p. 116], provides us with $\|\nabla x\|_{L^\infty} \le \gamma(\|u\|_U)$; that is, (55) holds.

It must be noted that, in the case of the p-Laplacian, the strong ellipticity of a holds for $p \ge 2$ and has the form

$$(a(r_1) - a(r_2))(r_1 - r_2) \ge K||r_1 - r_2||^p_{\mathbb{R}^N},\ r_1, r_2 \in \mathbb{R}^N,\ K > 0.$$

This does not affect the computations in all of our previous results, since our state equation takes place in $L^\infty(\Omega)$; the only difference is that sometimes we replace the power 2 by p.

Therefore if (x^*u^*) is an optimal pair for our problem, then there exist $\pi \in H^1_0(\Omega) \cap H^2(\Omega)$, $\gamma, l, \eta \in \mathbb{L}^2(\Omega)$ such that

$$\partial g(x^*) + l + \gamma \ni 0,\ B^*\pi \in \partial h(u^*),\ l = -\text{div } \eta, \tag{68}$$

and a.e. ω in Ω

$$\gamma(\omega) \in \pi(\omega)\partial\beta(x^*(\omega))\ , \tag{69}$$

$$\eta \in \left((p-2)\|\nabla x^*\|^{p-4} \sum_{j=1}^N \frac{\partial x^*}{\partial \omega_i}\frac{\partial x^*}{\partial \omega_j}\frac{\partial \pi}{\partial \omega_j} + \|\nabla x^*\|^{p-2}\frac{\partial \pi}{\partial \omega_i} \right)^T_{1 \le i \le N}. \tag{70}$$

Example 3.5. Consider the following version of (P) where the state equation is a modified form of the minimal surface equation:

$$\text{(P)} \qquad \text{minimize } g(x) + h(u) \text{ on all } (x,u) \in H^1_0(\Omega) \times U, \text{ subject to}$$

$$\text{(NE)} \quad -\text{div}\left(\frac{\nabla x}{\sqrt{1 + \|\nabla x\|^2_{\mathbb{R}^N}}} + k\nabla x \right) + \beta(x) = Bu + f \text{ in } H^1_0(\Omega),$$

where $N \ge 1$, $k > 1$, and g, h, β, B satisfy the assumptions in Theorem 2.1.

The minimal surface equation $\text{div}\,(\frac{\nabla x}{\sqrt{1+\|\nabla x\|^2}}) = 0$ appears in the non-parametric plateau problem of finding a surface with a given boundary and least possible area. The variational interpretation of the modified minimal surface equation describes the solutions of (NE) as minimizers of the energy $E : H^1_0(\Omega) \to \mathbb{R}$ given by

$$\begin{aligned} E(x) = \int_\Omega \sqrt{1 + \|\nabla x\|^2_{\mathbb{R}^N}}\, d\omega + \frac{k}{2}\int_\Omega (1 + \|\nabla x\|^2_{\mathbb{R}^N})d\omega \\ + \int_\Omega j(x)d\omega - \int_\Omega (Bu+f)x\, d\omega, \end{aligned}$$

$x \in H_0^1(\Omega)$. Here $A(x) = \int_\Omega \sqrt{1 + \|\nabla x\|^2_{\mathbb{R}^N}} d\omega$ represents the area of the surface S: $x_{N+1} = x(\omega)$, $\omega \in \Omega$, $\int_\Omega 1 + \|\nabla x\|^2_{\mathbb{R}^N} d\omega = \int_S \|\overrightarrow{n}\|_{\mathbb{R}^{N+1}} dS$, where $\overrightarrow{n} = (\nabla x, -1)$ is the gradient vector field on S, and $\beta = \partial j$.

In this case $a(r) = \frac{r}{\sqrt{1+\|r\|^2_{\mathbb{R}^N}}} + kr$, $r \in \mathbb{R}^N$, and by a simple verification involving the inequality $(r_1, r_2)_{\mathbb{R}^N} \leq \frac{1}{2}(\|r_1\|^2_{\mathbb{R}^N} + \|r_2\|^2_{\mathbb{R}^N})$ one gets that a is strongly monotone with constant $K = k - 1 > 0$ and Lipschitz continuous with Lipschitz constant $L = k+1$. The last assumption $\nabla^2 a \in L^\infty(\mathbb{R}^N; \mathbb{R}^{N^3})$ is checked by direct computation. Hence, by Theorem 2.1, for $(x^* u^*)$ an optimal pair (P) there exist $p \in H_0^1(\Omega) \cap H^2(\Omega)$, $\gamma, l, \eta \in \mathbb{L}^2(\Omega)$ such that

$$\partial g(x^*) + l + \gamma \ni 0,\ B^* p \in \partial h(u^*),\ l = -\text{div } \eta, \tag{71}$$

and a.e. ω in Ω

$$\gamma(\omega) \in p(\omega) \partial \beta(x^*(\omega)), \tag{72}$$

$$\eta \in \left(\left(k + \frac{1}{\sqrt{1 + \|\nabla x^*\|^2_{\mathbb{R}^N}}} \right) \frac{\partial p}{\partial \omega_i} - \sum_{j=1}^N \frac{\partial x^*}{\partial \omega_i} \frac{\partial x^*}{\partial \omega_j} \frac{\partial p}{\partial \omega_j} \frac{1}{(1 + \|\nabla x^*\|^2_{\mathbb{R}^N})^{3/2}} \right)^T_{1 \leq i \leq N}. \tag{73}$$

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Pure Appl. Math. (Amst.) 65, Academic Press, New York, 1975.

[2] N. Arada and J. P. Raymond, *State-constrained relaxed problems for semilinear elliptic equations*, J. Math. Anal. Appl., 223 (1998), pp. 248–271.

[3] V. Barbu, *Analysis and Control of Nonlinear Infinite Dimensional Systems*, Academic Press, Boston, MA, 1993.

[4] M. Bergounioux, *Optimal control of semilinear elliptic obstacle problems*, J. Nonlinear Convex Anal., 3 (2002), pp. 25–39.

[5] J. F. Bonnans and D. Tiba, *Pontryagin's principle in the control of semilinear elliptic variational inequalities*, Appl. Math. Optim., 23 (1991), pp. 299–312.

[6] E. Casas and L. A. Fernández, *Distributed control of systems governed by a general class of quasilinear elliptic equations*, J. Differential Equations, 104 (1993), pp. 20–47.

[7] E. Casas and J. M. Yong, *Maximum principle for state-constrained optimal control problems governed by quasilinear elliptic*, Differential Integral Equations, 8 (1995), pp. 1–18.

[8] E. Casas, F. Trölzsch, and A. Unger, *Second order sufficient optimality conditions for a nonlinear elliptic boundary control problem*, Z. Anal. Anwendungen, 15 (1996), pp. 687–707.

[9] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley and Sons, New York, 1983.

[10] R. Deville, G. Godefroy, and V. Zizler, *Smoothness and Renormings in Banach Spaces*, Pitman Monographs and Surveys in Pure and Applied Mathematics, 64 (1993).

[11] M. Dobrowolski and T. Staib, *Optimality conditions for state constrained nonlinear control problems*, Math. Methods Oper. Res., 42 (1995), pp. 129–160.

[12] L. C. Evans, *Partial Differential Equations*, Grad. Stud. Math. 19, American Mathematical Society, Providence, RI, 1998.

[13] L. Gasiński and N. S. Papageorgiou, *Nonsmooth Critical Point Theory and Nonlinear Boundary Value Problems*, Ser. Math. Anal. Appl. 8, Chapman & Hall/CRC, Boca Raton, FL, 2005.

[14] M. Marcus and V. J. Mizel, *Every superposition operator mapping one Sobolev space into another is continuous*, J. Funct. Anal., 33 (1979), pp. 217–229.

[15] C. Meyer and F. Trölzsch, *On an elliptic optimal control problem with pointwise mixed control-state constraints*, in Recent Advances in Optimization, Lecture Notes in Econom. and Math. Systems 563, Springer, Berlin, 2006, pp. 187–204.

[16] H. Sohr, *The Navier-Stokes Equations. An Elementary Functional Analytic Approach*, Birkhäuser Verlag, Basel, 2001.

[17] M. D. Voisei, *First-order necessary conditions of optimality for nonlinear programming problems associated with linear operators*, Adv. Math. Sci. Appl., 13 (2003), pp. 685–694.

[18] M. D. Voisei, *First-order necessary optimality conditions for nonlinear optimal control problems*, Panamer. Math. J., 14 (2004), pp. 1–44.
[19] M. D. Voisei, *First-order necessary conditions of optimality for strongly monotone nonlinear control problems*, J. Optim. Theory Appl., 116 (2003), pp. 421–436.
[20] M. D. Voisei, *First-Order Necessary Optimality Conditions for Nonlinear Optimal Control Problems*, Ph.D. thesis, Ohio University, 2004. Available online at www.ohiolink.edu/etd/view.cgi?ohiou1091111473.
[21] C. Zălinescu, *Convex Analysis in General Vector Spaces*, World Scientific, River Edge, NJ, 2002.

SIAM J. OPTIM.
Vol. 18, No. 4, pp. 1250–1265

PERTURBATION BOUNDS OF P-MATRIX LINEAR COMPLEMENTARITY PROBLEMS*

XIAOJUN CHEN† AND SHUHUANG XIANG‡

Abstract. We define a new fundamental constant associated with a P-matrix and show that this constant has various useful properties for the P-matrix linear complementarity problems (LCP). In particular, this constant is sharper than the Mathias–Pang constant in deriving perturbation bounds for the P-matrix LCP. Moreover, this new constant defines a measure of sensitivity of the solution of the P-matrix LCP. We examine how perturbations in the data affect the solution of the LCP and efficiency of Newton-type methods for solving the LCP.

Key words. perturbation bounds, sensitivity, linear complementarity problems

AMS subject classifications. 90C33, 65G20,65G50

DOI. 10.1137/060653019

1. Introduction.

The linear complementarity problem (LCP) is to find a vector $x \in R^n$ such that

$$Mx + q \geq 0, \quad x \geq 0, \quad x^T(Mx + q) = 0,$$

or to show that no such vector exists, where $M \in R^{n\times n}$ and $q \in R^n$. We denote this problem by LCP(M, q). A matrix M is called a P-matrix if all of its principal minors are positive, which is equivalent to

$$\max_{1\leq i\leq n} x_i(Mx)_i > 0 \quad \text{for all} \quad x \neq 0.$$

It is well known that M is a P-matrix if and only if the LCP(M, q) has a unique solution for any $q \in R^n$ [3]. Moreover, if M is a P-matrix, then there is a neighborhood $\mathcal{M}$ of M, such that all matrices in $\mathcal{M}$ are P-matrices. Hence, we can define a solution function $x(A, b) : \mathcal{M} \times R^n \to R^n_+$, where $x(A, b)$ is the solution of LCP(A, b) and $R^n_+ = \{x \in R^n \mid x \geq 0\}$.

In [12], Mathias and Pang introduced the following fundamental quantity associated with a P-matrix:

$$c(M) = \min_{\|x\|_\infty=1} \max_{1\leq i\leq n} \{x_i(Mx)_i\}.$$

This constant has often been used in error analysis of the LCP [2, 3]. In particular, the following lemma has been widely applied in perturbation bounds.

LEMMA 1.1 (see [3]). *Let $M \in R^{n\times n}$ be a P-matrix. The following statements hold:*

*Received by the editors February 26, 2006; accepted for publication (in revised form) April 6, 2007; published electronically October 17, 2007. This work is partly supported by a Grant-in-Aid from Japan Society for the Promotion of Science.

http://www.siam.org/journals/siopt/18-4/65301.html

†Department of Mathematical Sciences, Hirosaki University, Hirosaki 036-8561, Japan (chen@cc.hirosaki-u.ac.jp).

‡Department of Applied Mathematics and Software, Central South University, Changsha, Hunan 410083, China (xiangsh@mail.csu.edu.cn). The work of this author is also supported by the program of New Century Excellent Talents in University, Ministry of Education, People's Republic of China.

(i) *For any two vectors q and p in R^n,*

$$\|x(M,q) - x(M,p)\|_\infty \le \frac{1}{c(M)}\|q-p\|_\infty.$$

(ii) *For each vector $q \in R^n$, there exist a neighborhood $\mathcal{U}$ of the pair (M,q) and a constant $c_0 > 0$ such that for any $(A,b),(B,p) \in \mathcal{U}$, A, B are P-matrices and*

$$\|x(A,b) - x(B,p)\|_\infty \le c_0(\|A-B\|_\infty + \|b-p\|_\infty).$$

Lemma 1.1 shows that when M is a P-matrix, for each q, $x(A,b)$ is a locally Lipschitzian function of (A,b) in a neighborhood of (M,q), and $x(M,b)$ is a globally Lipschitzian function of b. This property plays a very important role in the study of the LCP and mathematical programs with LCP constraints [11]. However, the constant $c(M)$ is difficult to compute, and c_0 is not specified. It is hard to use this lemma for verifying the accuracy of a computed solution of the LCP when the data (M,q) contain errors.

In this paper, we introduce a new constant for a P-matrix,

$$\beta_p(M) = \max_{d\in[0,1]^n} \|(I - D + DM)^{-1}D\|_p,$$

where $D = \operatorname{diag}(d)$ with $0 \le d_i \le 1$, $i = 1,2,\ldots,n$, and $\|\cdot\|_p$ is the matrix norm induced by the vector norm for $p \ge 1$.

Using the constant $\beta_p(M)$, we give perturbation bounds for M being a P-matrix as follows:

$$\|x(M,q) - x(M,p)\|_p \le \beta_p(M)\|q-p\|_p, \tag{1.1}$$

$$\|x(A,b) - x(B,p)\|_p \le \frac{\beta_p(M)^2}{(1-\eta)^2}\|(-p)_+\|_p\|A-B\|_p + \frac{\beta_p(M)}{1-\eta}\|b-p\|_p, \tag{1.2}$$

and

$$\frac{\|x(M,q) - x(A,b)\|_p}{\|x(M,q)\|_p} \le \frac{2\epsilon}{1-\eta}\beta_p(M)\|M\|_p, \tag{1.3}$$

where $\eta \in [0,1)$ and $\epsilon > 0$ can be chosen, $A, B \in \mathcal{M} := \{A \mid \beta_p(M)\|M - A\|_p \le \eta\}$, and $\|q-b\|_p \le \epsilon\|(-q)_+\|_p$.

The constant $\beta_p(M)$ has the following interesting properties.

- If M is a P-matrix, then for $\|\cdot\|_\infty$,

$$\beta_\infty(M) \le \frac{1}{c(M)}. \tag{1.4}$$

- If M is an H-matrix with positive diagonals, then for $\|\cdot\|_p$ with any $p \ge 1$,

$$\beta_p(M) \le \|\tilde{M}^{-1}\|_p, \tag{1.5}$$

where $\tilde{M}$ is the comparison matrix of M, that is,

$$\tilde{M}_{ii} = M_{ii}, \qquad \tilde{M}_{ij} = -|M_{ij}| \quad \text{for} \quad i \ne j.$$

- If M is an M-matrix, then for $\|\cdot\|_p$ with any $p \geq 1$,

$$\beta_p(M) = \|M^{-1}\|_p. \tag{1.6}$$

- If M is a symmetric positive definite matrix, then for $\|\cdot\|_2$,

$$\beta_2(M) = \|M^{-1}\|_2. \tag{1.7}$$

Inequalities (1.1) and (1.4) show that the constant $\beta(M)$ derives a new perturbation bound which is sharper than the bound in (i) of Lemma 1.1 in $\|\cdot\|_\infty$. Furthermore, Example 2.1 shows that $\beta(M)$ can be much smaller than $c(M)^{-1}$ in some cases. Inequality (1.3) indicates that the constant $\beta(M)\|M\|$ is a measure of sensitivity of the solution $x(M,q)$ of the LCP(M,q). Moreover, from (1.3), (1.6), and (1.7), it is interesting to see that the measure is expressed in the terms of the condition number of M, that is,

$$\kappa_p(M) := \|M^{-1}\|_p\|M\|_p = \beta_p(M)\|M\|_p$$

for M being an M-matrix with $p \geq 1$ and a symmetric positive definite matrix with $p = 2$. Hence, it makes a connection between perturbation bounds of the LCP and perturbation bounds of the systems of linear equations in the Newton-type methods for solving the LCP. Using this connection, we investigate the efficiency of Newton-type methods for solving the following two systems:

$$r(x) := \min(x, Mx + q) = 0 \tag{1.8}$$

and

$$F(x,y) := \begin{pmatrix} Mx + q - y \\ \min(x,y) \end{pmatrix} = 0. \tag{1.9}$$

It is known that for the P-matrix LCP, the system of linear equations in Newton-type methods for solving (1.8) or (1.9) is mathematically well defined; that is, the generalized Jacobian matrices are nonsingular [5]. However, the matrices can be computationally ill-conditioned. A matrix A is said to be an ill-conditioned (well-conditioned) matrix if $\kappa_p(A)$ is large (small) [8]. The condition number $\kappa_p(A)$ is a measure of sensitivity of the system of linear equations $Ax = b$ when A is nonsingular. Hence, a linear system is called ill-conditioned (well-conditioned) if $\kappa_p(A)$ is large (small) [4]. From (1.3), (1.6), and (1.7), we find that $\beta_p(M)\|M\|_p$ is a measure of sensitivity of the LCP(M,q) when M is a P-matrix, and $\beta_p(M)\|M\|_p = \kappa_p(M)$ when M is an M-matrix or a symmetric positive definite matrix. Moreover, we show that for the M-matrix LCP, the systems of linear equations in the Newton-type methods for solving (1.8) are well-conditioned if and only if the condition number $\kappa_p(M)$ is small. However, the system of linear equations in Newton-type methods for solving (1.9) can be ill-conditioned even when $\kappa_p(M)$ is small.

A word about our notation: For a vector $q \in R^n$, $q_+ = \max(0,q)$. Let $N = \{1,2,\ldots,n\}$. Let e be the vector whose elements are all 1. A matrix $A \in R^{n\times n}$ is called an M-matrix if $A^{-1} \geq 0$ and $A_{ij} \leq 0$ $(i \neq j)$ for $i,j \in N$; A is called an H-matrix if its comparison matrix is an M-matrix.

In the rest of this paper, we use $\beta(\cdot)$, $\|\cdot\|$, and $\kappa(\cdot)$ to represent $\beta_p(\cdot)$, $\|\cdot\|_p$, and $\kappa_p(\cdot)$ with any $p \geq 1$, respectively.

2. A new constant for the P-matrix LCP. In this section we introduce a new Lipschitz constant for the P-matrix LCP based on the observation that for any $x, x^*, y, y^* \in R^n$,

$$\min(x_i, y_i) - \min(x_i^*, y_i^*) = (1 - d_i)(x_i - x_i^*) + d_i(y_i - y_i^*), \qquad i \in N, \tag{2.1}$$

where

$$d_i = \begin{cases} 0 & \text{if } y_i \geq x_i,\ y_i^* \geq x_i^*, \\ 1 & \text{if } y_i \leq x_i,\ y_i^* \leq x_i^*, \\ \dfrac{\min(x_i, y_i) - \min(x_i^*, y_i^*) + x_i^* - x_i}{y_i - y_i^* + x_i^* - x_i} & \text{otherwise.} \end{cases}$$

It is easy to see that $d_i \in [0, 1]$. Set $x = x(A, q)$, $x^* = x(B, p)$, $y = Ax(A, q) + q$, and $y^* = Bx(B, p) + p$ in (2.1). We obtain

$$0 = (I - D)(x(A, q) - x(B, p)) + D(Ax(A, q) + q - Bx(B, p) - p),$$

which implies

$$(I - D + DA)(x(B, p) - x(A, q)) = D(A - B)x(B, p) + D(q - p). \tag{2.2}$$

Here D is a diagonal matrix whose diagonal elements are $d = (d_1, d_2, \ldots, d_n) \in [0, 1]^n$.

LEMMA 2.1 (Gabriel and Moré [7]). *A is a P-matrix if and only if $I - D + DA$ is nonsingular for any diagonal matrix $D = \operatorname{diag}(d)$ with $0 \leq d_i \leq 1$.*

For M being a P-matrix, we introduce the following constant:

$$\beta(M) = \max_{d \in [0,1]^n} \|(I - D + DM)^{-1} D\|.$$

From Lemma 2.1 and (2.2), we have

$$\|x(B, p) - x(A, q)\| \leq \beta(A)\|(A - B)x(B, p) + q - p\| \tag{2.3}$$

provided A is a P-matrix. In the following, we compare $\beta(M)$ with $c(M)^{-1}$ in $\|\cdot\|_\infty$ and give a simple version of $\beta(M)$ for M being an M-matrix, a symmetric positive definite matrix, and a positive definite matrix.

THEOREM 2.2. *Let M be a P-matrix. Then*

$$\beta_\infty(M) := \max_{d \in [0,1]^n} \|(I - D + DM)^{-1} D\|_\infty \leq \frac{1}{c(M)}.$$

Proof. We first prove that for any nonsingular diagonal matrix $D = \operatorname{diag}(d)$ with $d \in (0, 1]^n$,

$$\|(I - D + DM)^{-1} D\|_\infty \leq \frac{1}{c(M)}.$$

Let $x \in R^n$ with $\|x\|_\infty = 1$ such that $\|(I-D+DM)^{-1}D\|_\infty = \|(I-D+DM)^{-1}Dx\|_\infty$ and define $y = (I-D+DM)^{-1}Dx$. Then $Dx = (I-D+DM)y$, $My = x+y-D^{-1}y$. By the definition of $c(M)$, we have

$$0 < c(M)\|y\|_\infty^2 \leq \max_i y_i(My)_i = \max_i y_i \left(x_i + y_i - \frac{y_i}{d_i} \right).$$

Note that $f(t) = a(b + a - \frac{a}{t})$ is monotonically increasing for $t > 0$, where a, b are constants. Therefore, we deduce

$$y_i \left(x_i + y_i - \frac{y_i}{d_i} \right) \leq y_i x_i \leq \|y\|_\infty \|x\|_\infty = \|y\|_\infty,$$

which implies

$$0 < c(M)\|y\|_\infty^2 \leq \|y\|_\infty \quad \text{and} \quad \|(I - D + DM)^{-1} D\|_\infty = \|y\|_\infty \leq \frac{1}{c(M)}.$$

Now we consider $d \in [0,1]^n$. Let $d_\epsilon = \min(d + \epsilon e, e)$, where $\epsilon \in (0,1]$. Then, we have

$$\|(I - D + DM)^{-1} D\|_\infty = \lim_{\epsilon \downarrow 0} \|(I - D_\epsilon + D_\epsilon M)^{-1} D_\epsilon\|_\infty \leq \frac{1}{c(M)}. \qquad \square$$

It is known that an H-matrix with positive diagonals is a P-matrix, and a positive definite matrix is a P-matrix [3]. Now, we consider the two subclasses of P-matrix.

LEMMA 2.3 (see [3]). *If M is an M-matrix, then $I - D + DM$ is an M-matrix for $d \in [0,1]^n$.*

LEMMA 2.4. *Let A be an H-matrix with positive diagonals, and let $\tilde{A}$ be the comparison matrix of A. Then the following statements hold:*

(i) $|A^{-1}| \leq \tilde{A}^{-1}$.

(ii) *For $B \in R^{n \times n}$ with $\|B\|_\infty \|\tilde{A}^{-1}\|_\infty < 1$, $A + B$ is an H-matrix with positive diagonals.*

Proof. (i) See problem 31 in [10, page 131].

(ii) Let $x = \tilde{A}^{-1} e$. Since $\tilde{A}^{-1} \geq 0$, $x > 0$ and $\|x\|_\infty = \|\tilde{A}^{-1}\|_\infty$. Moreover, from $\tilde{A}x = e$, we have

$$a_{ii} x_i = 1 + \sum_{j \neq i} |a_{ij}| x_j \qquad \text{for } i \in N.$$

By $\|x\|_\infty \|B\|_\infty < 1$ and $\|B\|_\infty = \||B|e\|_\infty$, we get $\|x\|_\infty |B| e < e$. Hence for all $i \in N$,

$$a_{ii} x_i > \sum_{j=1}^{n} |b_{ij}| \|x\|_\infty + \sum_{j \neq i} |a_{ij}| x_j \geq \sum_{j \neq i} (|a_{ij}| + |b_{ij}|) x_j + |b_{ii}| x_i,$$

and

$$(a_{ii} + b_{ii}) x_i \geq (a_{ii} - |b_{ii}|) x_i > \sum_{j \neq i} (|a_{ij}| + |b_{ij}|) x_j.$$

By I_{27} of Theorem 2.3 in [1, Chap. 6], this implies that the comparison matrix of $A + B$ is an M-matrix. Hence $A + B$ is an H-matrix with positive diagonals. $\square$

THEOREM 2.5. *Let M be an H-matrix with positive diagonals. Then*

$$\beta(M) \leq \|\tilde{M}^{-1}\|,$$

where $\tilde{M}$ is the comparison matrix of M. In particular, if M is an M-matrix, then the equality holds with $M = \tilde{M}$.

Proof. First we will show that if M is an M-matrix, then

$$\beta(M) = \|M^{-1}\|.$$

Since for any $d \in (0,1]^n$, by Lemma 2.3,

$$(DM)^{-1} - (I - D + DM)^{-1} = (DM)^{-1}(I - D)(I - D + DM)^{-1} \geq 0,$$

we have

$$(DM)^{-1}D - (I - D + DM)^{-1}D = M^{-1} - (I - D + DM)^{-1}D \geq 0.$$

Note that for any matrices A and B, $|A| \leq B$ implies $\|A\| \leq \|B\|$. Hence the following inequalities hold:

$$M^{-1} \geq (I - D + DM)^{-1}D \geq 0, \qquad \|M^{-1}\| \geq \|(I - D + DM)^{-1}D\|.$$

Let $d_\epsilon = \min(d + \epsilon e, e)$, where $\epsilon \in (0,1]$. Then, we have

$$\beta(M) = \max_{d \in [0,1]^n} \lim_{\epsilon \downarrow 0} \|(I - D_\epsilon + D_\epsilon M)^{-1} D_\epsilon\| \leq \|M^{-1}\|.$$

It is obvious that $\beta(M) \geq \|M^{-1}\|$ as $e \in [0,1]^n$. Therefore, we obtain $\beta(M) = \|M^{-1}\|$.

For M being an H-matrix, $\tilde{M}$ is an M-matrix. From (i) of Lemma 2.4, we have

$$|(I - D + DM)^{-1}| \leq (I - D + D\tilde{M})^{-1}.$$

Hence, we obtain

$$\beta(M) = \max_{d \in [0,1]^n} \|(I - D + DM)^{-1}D\| \leq \max_{d \in [0,1]^n} \|(I - D + D\tilde{M})^{-1}D\| \leq \|\tilde{M}^{-1}\|. \qquad \square$$

LEMMA 2.6 (see [9]). *Let A and B be symmetric positive definite matrices.*

(i) *$B - A$ is positive semidefinite if and only if $A^{-1} - B^{-1}$ is positive semidefinite.*

(ii) *If $B - A$ is positive semidefinite, then $\lambda_i(B) \geq \lambda_i(A)$, where $\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_n(A)$ and $\lambda_1(B) \geq \lambda_2(B) \geq \cdots \geq \lambda_n(B)$ are eigenvalues of A and B, respectively.*

THEOREM 2.7. *Let M be a symmetric positive definite matrix. Then*

$$\beta_2(M) := \max_{d \in [0,1]^n} \|(I - D + DM)^{-1}D\|_2 = \|M^{-1}\|_2.$$

Proof. It is obvious that $\beta_2(M) \geq \|M^{-1}\|_2$. Now we show $\beta_2(M) \leq \|M^{-1}\|_2$. For any nonsingular diagonal matrix $D = \operatorname{diag}(d)$ with $d \in (0,1]^n$, $M + D^{-1}(I - D)$ is positive definite. By (i) of Lemma 2.6, $M^{-1} - (M + D^{-1}(I - D))^{-1}$ is positive semidefinite. By (ii) of Lemma 2.6, we have

$$\|(M + D^{-1}(I - D))^{-1}\|_2 = \|(I - D + DM)^{-1}D\|_2 \leq \|M^{-1}\|_2.$$

Since the largest eigenvalue is a continuous function of elements of the matrix, we have

$$\beta_2(M) = \max_{d \in [0,1]^n} \lim_{\epsilon \downarrow 0} \|(I - D_\epsilon + D_\epsilon M)^{-1} D_\epsilon\|_2 \leq \|M^{-1}\|_2,$$

where $D_\epsilon = \operatorname{diag}(\min(d + \epsilon e, e))$. $\square$

In comparison to Lemma 1.1, the following theorem gives sharp perturbation error estimates for the P-matrix LCP.

THEOREM 2.8. *Let $M \in R^{n \times n}$ be a P-matrix. Then the following statements hold:*

(i) *For any two vectors q and p in R^n,*

$$\|x(M,q) - x(M,p)\| \le \beta(M)\|q-p\|.$$

(ii) *Every matrix $A \in \mathcal{M} := \{A \mid \beta(M)\|M-A\| \le \eta < 1\}$ is a P-matrix. Let*

$$\alpha(M) = \frac{1}{1-\eta}\beta(M).$$

Then for any $A, B \in \mathcal{M}$ and $q, p \in R^n$

$$\|x(A,q) - x(B,p)\| \le \alpha(M)^2\|(-p)_+\|\|A-B\| + \alpha(M)\|q-p\|.$$

Proof. (i) This part of the proof follows directly from (2.3) by setting $M = A = B$.
(ii) For every $A \in \mathcal{M}$, since $\|(I-D+DM)^{-1}D(A-M)\| \le \beta(M)\|M-A\| \le \eta < 1$,

$$(I - D + DA) = (I - D + DM)(I + (I - D + DM)^{-1}D(A - M))$$

is nonsingular for any diagonal matrix $D = \mathrm{diag}(d)$ with $0 \le d_i \le 1$. By Lemma 2.1, A is a P-matrix. Moreover, from

$$(I - D + DA)^{-1}D = (I + (I - D + DM)^{-1}D(A - M))^{-1}(I - D + DM)^{-1}D$$

and

$$\|(I + (I - D + DM)^{-1}D(A - M))^{-1}\| \le \frac{1}{1 - \beta(M)\|A - M\|} \le \frac{1}{1-\eta},$$

we find $\beta(A) \le \alpha(M)$.

Since matrices $A, B \in \mathcal{M}$ are P-matrices, using (2.3) yields

$$\|x(A,q) - x(B,p)\| \le \beta(A)\left(\|A-B\|\|x(B,p)\| + \|q-p\|\right). \tag{2.4}$$

Notice that 0 is the solution of LCP(B, p_+). Using (2.3) again, we get

$$\|x(B,p)\| \le \beta(B)\|(-p)_+\|. \tag{2.5}$$

Applying $\beta(A) \le \alpha(M)$ and $\beta(B) \le \alpha(M)$ to (2.4) and (2.5), respectively, we obtain the desired bounds in (ii). ∎

From Theorems 2.5 and 2.7, the Lipschitz constants $\beta(M)$ and $\alpha(M)$ can be estimated by matrix norms if M is an H-matrix with positive diagonals or a symmetric positive definite matrix. In particular, from Lemma 2.4, Theorem 2.5, and Theorem 2.7, we have the following two corollaries.

COROLLARY 2.9. *Let $M \in R^{n\times n}$ be an H-matrix with positive diagonals. Then the following statements hold:*
(i) *For any two vectors q and p in R^n,*

$$\|x(M,q) - x(M,p)\|_\infty \le \|\tilde{M}^{-1}\|_\infty\|q-p\|_\infty.$$

(ii) *Every matrix $A \in \mathcal{M}_\infty := \{A \mid \|\tilde{M}^{-1}\|_\infty\|M-A\|_\infty \le \eta < 1\}$ is an H-matrix with positive diagonals. Let*

$$\alpha_\infty(M) = \frac{1}{1-\eta}\|\tilde{M}^{-1}\|_\infty.$$

Then for any $A, B \in \mathcal{M}_\infty$ and $q, p \in R^n$

$$\|x(A,q) - x(B,p)\|_\infty \le \alpha_\infty(M)^2\|(-p)_+\|_\infty\|A-B\|_\infty + \alpha_\infty(M)\|q-p\|_\infty.$$

COROLLARY 2.10. *Let $M \in R^{n\times n}$ be a symmetric positive definite matrix. Then the following statements hold:*
(i) *For any two vectors q and p in R^n,*

$$\|x(M,q) - x(M,p)\|_2 \le \|M^{-1}\|_2 \|q-p\|_2.$$

(ii) *Every matrix $A \in \mathcal{M}_2 := \{A \mid \|M^{-1}\|_2\|M-A\|_2 \le \eta < 1\}$ is a P-matrix. Let*

$$\alpha_2(M) = \frac{1}{1-\eta}\|M^{-1}\|_2.$$

Then for any $A, B \in \mathcal{M}_2$ and $q, p \in R^n$

$$\|x(A,q) - x(B,p)\|_2 \le \alpha_2(M)^2\|(-p)_+\|_2\|A-B\|_2 + \alpha_2(M)\|q-p\|_2.$$

A matrix A is called positive definite if

$$x^T Ax > 0, \qquad 0 \ne x \in R^n.$$

Since $x^T Ax = x^T \frac{A+A^T}{2} x$, A is positive definite if and only if $\frac{A+A^T}{2}$ is symmetric positive definite. Note that a positive definite matrix is not necessarily symmetric. Such asymmetric matrices frequently appear in the context of the LCP.

Combining the ideas of Mathias and Pang [12] and Corollary 2.10, we present perturbation bounds for the positive definite matrix LCP.

THEOREM 2.11. *Let $M \in R^{n\times n}$ be a positive definite matrix. Then the following statements hold:*
(i) *For any two vectors q and p in R^n,*

$$\|x(M,q) - x(M,p)\|_2 \le \left\| \left(\frac{M+M^T}{2} \right)^{-1} \right\|_2 \|q-p\|_2.$$

(ii) *Every matrix $A \in \mathcal{M}_2 := \{A \mid \|(\frac{M+M^T}{2})^{-1}\|_2\|M-A\|_2 \le \eta < 1\}$ is positive definite. Let*

$$\alpha_2(M) = \frac{1}{1-\eta} \left\| \left(\frac{M+M^T}{2} \right)^{-1} \right\|_2.$$

Then for any $A, B \in \mathcal{M}_2$ and $q, p \in R^n$

$$\|x(A,q) - x(B,p)\|_2 \le \alpha_2(M)^2\|(-p)_+\|_2\|A-B\|_2 + \alpha_2(M)\|q-p\|_2.$$

Proof. We first show that

$$\|x(A,q) - x(B,p)\|_2 \le \left\| \left(\frac{A+A^T}{2} \right)^{-1} \right\|_2 (\|A-B\|_2\|x(B,p)\|_2 + \|p-q\|_2) \tag{2.6}$$

holds if A is a positive definite matrix and the LCP(B,p) has a solution $x(B,p)$.

Since $x(A,q)$ and $x(B,p)$ are solutions of LCP(A,q) and LCP(B,p), respectively, we have

$$\begin{aligned} 0 &\ge (x(A,q) - x(B,p))^T (Ax(A,q) + q - Bx(B,p) - p) \\ &= (x(A,q) - x(B,p))^T (A(x(A,q) - x(B,p)) + (A-B)x(B,p) + q - p), \end{aligned}$$

which implies

$$\begin{aligned}
&(x(A,q)-x(B,p))^T((B-A)x(B,p)+p-q)\\
&\geq (x(A,q)-x(B,p))^T A(x(A,q)-x(B,p))\\
&= (x(A,q)-x(B,p))^T \frac{A+A^T}{2}(x(A,q)-x(B,p))\\
&\geq \frac{1}{\|(\frac{A+A^T}{2})^{-1}\|_2}\|x(A,q)-x(B,p)\|_2^2.
\end{aligned}$$

Using the Cauchy–Schwarz inequality, we get (2.6).

(i) Set $A=B=M$ in (2.6); we get the desired bound.

(ii) Note that for any matrix C, $\|C\|_2=\|C^T\|_2$. For any $x\in R^n$ with $x\neq 0$, we have

$$\begin{aligned}
x^TAx &= x^T\frac{M+M^T}{2}x + x^T\left(\frac{A+A^T}{2}-\frac{M+M^T}{2}\right)x\\
&\geq x^T\frac{M+M^T}{2}x - \left(\left\|\frac{A-M}{2}\right\|_2+\left\|\frac{A^T-M^T}{2}\right\|_2\right)\|x\|_2^2\\
&\geq \left(\left\|\left(\frac{M+M^T}{2}\right)^{-1}\right\|_2\right)^{-1}\|x\|_2^2 - \|A-M\|_2\|x\|_2^2\\
&\geq \left(\left\|\left(\frac{M+M^T}{2}\right)^{-1}\right\|_2\right)^{-1}\left(1-\left\|\left(\frac{M+M^T}{2}\right)^{-1}\right\|_2\|M-A\|_2\right)\|x\|_2^2.
\end{aligned}$$

Hence for any $A\in\mathcal{M}_2$, $x^TAx>0$, and thus A is positive definite. Moreover, from

$$\left(\frac{A+A^T}{2}\right)^{-1}=\left(I+\left(\frac{M+M^T}{2}\right)^{-1}\left(\frac{A+A^T}{2}-\frac{M+M^T}{2}\right)\right)^{-1}\left(\frac{M+M^T}{2}\right)^{-1}$$

and

$$\left\|\frac{A+A^T}{2}-\frac{M+M^T}{2}\right\|_2\leq\frac{1}{2}(\|A-M\|_2+\|A^T-M^T\|_2)=\|M-A\|_2$$

we have

$$\begin{aligned}
\left\|\left(\frac{A+A^T}{2}\right)^{-1}\right\|_2 &\leq \frac{1}{1-\|(\frac{M+M^T}{2})^{-1}\|_2\|\frac{A+A^T}{2}-\frac{M+M^T}{2}\|_2}\left\|\left(\frac{M+M^T}{2}\right)^{-1}\right\|_2\\
&\leq \alpha_2(M).
\end{aligned}$$

Similarly, for $B\in\mathcal{M}_2$, $\|(\frac{B+B^T}{2})^{-1}\|_2\leq\alpha_2(M)$. Notice that 0 is the solution of LCP(B,p_+). Setting $A=B$ and $q=p_+$ in (2.6), we get

$$\|x(B,p)\|_2\leq\left\|\left(\frac{B+B^T}{2}\right)^{-1}\right\|_2\|(-p)_+\|_2.$$

Using these inequalities with (2.6), we obtain the perturbation bound in (ii). ∎

Example 2.1. Theorem 2.2 shows that for every P-matrix, $\beta_\infty(M) \le c(M)^{-1}$. Now we show that $\beta_\infty(M)$ can be much smaller than $c(M)^{-1}$ in some cases. Consider

$$M = \begin{pmatrix} 1 & -t \\ 0 & t \end{pmatrix}.$$

For $t \ge 1$, M is an M-matrix. By Theorem 2.5, $\beta_\infty(M) = \|M^{-1}\|_\infty = 2$. For $\bar{x} = (1, t^{-1})$, we have

$$c(M) \le \max_{i \in N} \bar{x}_i (M\bar{x})_i = \frac{1}{t}.$$

Hence, $c(M)^{-1} \ge t \to \infty$ as $t \to \infty$.

3. Relative perturbation bounds for the LCP. Using the results in the last section, we derive relative perturbation bounds expressed in the term of $\beta(M)\|M\|$.

THEOREM 3.1. *Suppose*

$$\begin{aligned} \min(x, Mx+q) &= 0, \qquad M \in R^{n\times n}, \ \ 0 \ne (-q)_+ \in R^n, \\ \min(y, (M+\triangle M)y + q + \triangle q) &= 0, \qquad \triangle M \in R^{n\times n}, \ \ \triangle q \in R^n, \end{aligned}$$

with

$$\|\triangle M\| \le \epsilon \|M\|, \qquad \|\triangle q\| \le \epsilon \|(-q)_+\|.$$

If M is a P-matrix and $\epsilon\beta(M)\|M\| = \eta < 1$, then $M + \triangle M$ is a P-matrix and

$$\frac{\|y - x\|}{\|x\|} \le \frac{2\epsilon}{1-\eta}\beta(M)\|M\|.$$

Proof. First we observe that x is a solution of LCP(M, q) and y is a solution of LCP$(M + \triangle M, q + \triangle q)$. Then following the proof of (ii) of Theorem 2.8, we obtain that $M + \triangle M$ is a P-matrix and

$$\beta(M + \triangle M) \le \frac{1}{1-\eta}\beta(M),$$

which, together with (2.3), gives

$$\|x - y\| \le \frac{1}{1-\eta}\beta(M)(\|\triangle M\|\|x\| + \|\triangle q\|). \tag{3.1}$$

From $Mx + q \ge 0$, we deduce $(-q)_+ \le (Mx)_+ \le |Mx|$. This implies $\|(-q)_+\| \le \|Mx\| \le \|M\|\|x\|$. Hence, we have

$$\|x\| \ge \frac{1}{\|M\|}\|(-q)_+\| > 0. \tag{3.2}$$

Combining (3.1) and (3.2), we obtain the desired bounds

$$\frac{\|y - x\|}{\|x\|} \le \frac{1}{1-\eta}\beta(M)\left(\|\triangle M\| + \frac{\|\triangle q\|}{\|x\|}\right) \le \frac{2\epsilon}{1-\eta}\beta(M)\|M\|. \qquad \blacksquare$$

Theorem 3.1 indicates that $\beta(M)\|M\|$ is a measure of sensitivity of the solution of the LCP(M,q) for M being a P-matrix. Moreover, Theorem 3.1 with Corollary 2.9, Corollary 2.10, and Theorem 2.11 gives $\beta(M)\|M\|$ in the term of condition number for the H-matrix LCP, symmetric positive definite LCP, and positive definite LCP.

COROLLARY 3.2. *Suppose*

$$\begin{aligned} \min(x, Mx+q) &= 0, \qquad M \in R^{n\times n}, \ \ 0 \neq (-q)_+ \in R^n, \\ \min(y, (M+\triangle M)y + q + \triangle q) &= 0, \qquad \triangle M \in R^{n\times n}, \ \ \triangle q \in R^n. \end{aligned}$$

(i) *If M is an H-matrix with positive diagonals, $\epsilon\kappa_\infty(\tilde{M}) = \eta < 1$, and*

$$\|\triangle M\|_\infty \leq \epsilon \|\tilde{M}\|_\infty, \qquad \|\triangle q\|_\infty \leq \epsilon \|(-q)_+\|_\infty,$$

then $M + \triangle M$ is an H-matrix with positive diagonals, and

$$\frac{\|y-x\|_\infty}{\|x\|_\infty} \leq \frac{2\epsilon}{1-\eta}\kappa_\infty(\tilde{M}).$$

(ii) *If M is a symmetric positive definite matrix, $\epsilon\kappa_2(M) = \eta < 1$, and*

$$\|\triangle M\|_2 \leq \epsilon \|M\|_2, \qquad \|\triangle q\|_2 \leq \epsilon \|(-q)_+\|_2,$$

then $M + \triangle M$ is a P-matrix, and

$$\frac{\|y-x\|_2}{\|x\|_2} \leq \frac{2\epsilon}{1-\eta}\kappa_2(M).$$

(iii) *If M is a positive definite matrix, $\epsilon\kappa_2(\frac{M+M^T}{2}) = \eta < 1$, and*

$$\|\triangle M\|_2 \leq \epsilon \left\| \frac{M+M^T}{2} \right\|_2, \qquad \|\triangle q\|_2 \leq \epsilon \|(-q)_+\|_2 \frac{\|M+M^T\|_2}{2\|M\|_2},$$

then $M + \triangle M$ is a positive definite matrix, and

$$\frac{\|x-y\|_2}{\|x\|_2} \leq \frac{2\epsilon}{1-\eta}\kappa_2\left(\frac{M+M^T}{2}\right).$$

Remark. Note that $\|(-q)_+\| \leq \|q\|$. If $Mx + q = 0$, then (i) of Corollary 3.2 for M being an M-matrix and (ii) of Corollary 3.2 reduce to the perturbation bounds for the system of linear equations [8].

For the H-matrix LCP, componentwise perturbation bounds based on the Skeel condition number $\||\tilde{M}^{-1}||\tilde{M}|\|_\infty$ can be represented as follows.

THEOREM 3.3. *Suppose*

$$\begin{aligned} \min(x, Mx+q) &= 0, \qquad M \in R^{n\times n}, \ \ 0 \neq (-q)_+ \in R^n, \\ \min(y, (M+\triangle M)y + q + \triangle q) &= 0, \qquad \triangle M \in R^{n\times n}, \ \ \triangle q \in R^n, \end{aligned}$$

with

$$|\triangle M| \leq \epsilon |M|, \qquad |\triangle q| \leq \epsilon (-q)_+. \tag{3.3}$$

If M is an H-matrix with positive diagonals and $\epsilon\kappa_\infty(\tilde{M}) = \eta < 1$, then $M + \triangle M$ is an H-matrix with positive diagonals, and

$$\frac{\|y - x\|_\infty}{\|x\|_\infty} \leq \frac{2\epsilon}{1-\eta}\|\tilde{M}^{-1}|\tilde{M}|\|_\infty. \tag{3.4}$$

Proof. From (3.3), we have

$$\|\triangle M\|_\infty \leq \epsilon\|\tilde{M}\|_\infty, \quad \text{and} \quad \|\triangle q\|_\infty \leq \epsilon\|(-q)_+\|_\infty \leq \epsilon\|M\|_\infty\|x\|_\infty,$$

where the last inequality uses $(-q)_+ \leq (Mx)_+ \leq |M|x$.

According to Corollary 3.2, $M + \triangle M$ is an H-matrix with positive diagonals. Moreover, the equality (2.2) gives

$$(I - D + DM)(y - x) = D\triangle My + D\triangle q \tag{3.5}$$

for some diagonal matrix $D = \mathrm{diag}(d)$ with $d \in [0,1]^n$.

Following the proof of Theorem 2.5, by Lemma 2.4, we get

$$\begin{aligned}
|y - x| &\leq |(I - D + DM)^{-1}D|(|\triangle M|y + |\triangle q|) \\
&\leq |(I - D + D\tilde{M})^{-1}D|(|\triangle M|y + |\triangle q|) \\
&\leq \tilde{M}^{-1}(|\triangle M|y + |\triangle q|) \\
&\leq \epsilon\tilde{M}^{-1}(|M|y + |M|x).
\end{aligned}$$

Therefore, we find

$$\|y - x\|_\infty \leq \epsilon\|\tilde{M}^{-1}|M|\|_\infty(\|y\|_\infty + \|x\|_\infty). \tag{3.6}$$

Furthermore, from (3.5), we obtain

$$y - ((I - D + DM)^{-1}D\triangle M)y = x + (I - D + DM)^{-1}D\triangle q.$$

Hence, it holds that

$$\begin{aligned}
(1 - \epsilon\|\tilde{M}^{-1}\|_\infty\|M\|_\infty)\|y\|_\infty &\leq (1 - \|(I - D + DM)^{-1}D\|_\infty\|\triangle M\|_\infty)\|y\|_\infty \\
&\leq \|y - (I - D + DM)^{-1}D\triangle My\|_\infty \\
&\leq \|x\|_\infty + \|(I - D + DM)^{-1}D\|_\infty\|\triangle q\|_\infty \\
&\leq (1 + \epsilon\|\tilde{M}^{-1}\|_\infty\|M\|_\infty)\|x\|_\infty.
\end{aligned}$$

This implies

$$\|y\|_\infty \leq \frac{1+\eta}{1-\eta}\|x\|_\infty. \tag{3.7}$$

Combining (3.6) and (3.7), we obtain the desired bounds (3.4). ∎

4. Newton-type methods. In the last two sections, we have given perturbation bounds for the LCP in the term of $\beta(M)$. In this section, we use the perturbation bounds to analyze the efficiency of Newton-type methods for solving the LCP based on the systems (1.8) and (1.9).

Many semismooth Newton methods, smoothing Newton methods, and path-following interior point methods [5] solve a system of linear equations in each iteration,

$$(I - D_k + D_k M)(x - x^k) = -r(x^k), \tag{4.1}$$

or

$$\begin{pmatrix} M & -I \\ I - D_k & D_k \end{pmatrix} \begin{pmatrix} x - x^k \\ y - y^k \end{pmatrix} = -F(x^k, y^k), \tag{4.2}$$

where D_k is a diagonal matrix whose diagonal elements are in $[0, 1]$.

Sensitivity of (4.1) and (4.2) will affect implementation of the methods and reliability of the computed solution. From the analysis of Dennis and Schnabel [4], if the condition number of the coefficient matrix of the linear equations is larger than $(macheps)^{-1/2}$, the numerical solution may not be trustworthy. Here *macheps* is computer precision. The linear systems (4.1) and (4.2) have the following relation regarding to the condition numbers.

PROPOSITION 4.1. *For any diagonal matrix* $D = \mathrm{diag}(d)$ *with* $0 \le d_i \le 1$, $i = 1, 2, \ldots, n$, *the following inequalities hold:*

$$\kappa_\infty \begin{pmatrix} M & -I \\ I - D & D \end{pmatrix} \ge \kappa_\infty (I - D + DM) \tag{4.3}$$

and

$$\kappa \begin{pmatrix} M & -I \\ I - D & D \end{pmatrix} \ge \frac{1}{2} \kappa (I - D + DM). \tag{4.4}$$

Proof. First, we observe

$$\left\| \begin{pmatrix} M & -I \\ I - D & D \end{pmatrix} \right\|_\infty \ge 1 + \|M\|_\infty \ge \max(1, \|M\|_\infty) \ge \|I - D + DM\|_\infty$$

and

$$\left\| \begin{pmatrix} M & -I \\ I - D & D \end{pmatrix} \right\| \ge \max(1, \|M\|) \ge \frac{\max(1, \|M\|)}{1 + \|M\|} \|I - D + DM\| \ge \frac{1}{2} \|I - D + DM\|.$$

Next, we consider the inverses. From

$$\begin{pmatrix} I & 0 \\ D & I \end{pmatrix} \begin{pmatrix} M & -I \\ I - D & D \end{pmatrix} = \begin{pmatrix} M & -I \\ I - D + DM & 0 \end{pmatrix}$$

and

$$\begin{pmatrix} M & -I \\ I - D + DM & 0 \end{pmatrix}^{-1} = \begin{pmatrix} 0 & (I - D + DM)^{-1} \\ -I & M(I - D + DM)^{-1} \end{pmatrix},$$

we find the inverse

$$\begin{aligned} \begin{pmatrix} M & -I \\ I - D & D \end{pmatrix}^{-1} &= \begin{pmatrix} 0 & (I - D + DM)^{-1} \\ -I & M(I - D + DM)^{-1} \end{pmatrix} \begin{pmatrix} I & 0 \\ D & I \end{pmatrix} \\ &= \begin{pmatrix} (I - D + DM)^{-1} D & (I - D + DM)^{-1} \\ M(I - D + DM)^{-1} D - I & M(I - D + DM)^{-1} \end{pmatrix}. \end{aligned}$$

Therefore, we have

$$\left\| \begin{pmatrix} M & -I \\ I-D & D \end{pmatrix}^{-1} \right\| \geq \|(I-D+DM)^{-1}\|.$$

By the definition of the condition number, (4.3) and (4.4) hold. □

Since D_k in the coefficient matrices of (4.1) and (4.2) changes at each step, we consider the worst case

$$K(M) := \max_{d\in[0,1]^n} \|(I-D+DM)^{-1}\|\|I-D+DM\|$$

and

$$\hat{K}(M) := \max_{d\in[0,1]^n} \left\| \begin{pmatrix} M & -I \\ D & I-D \end{pmatrix}^{-1} \right\| \left\| \begin{pmatrix} M & -I \\ D & I-D \end{pmatrix} \right\|.$$

From Proposition 4.1, we have

$$\hat{K}_\infty(M) \geq K_\infty(M),$$

which implies that if (4.2) is well-conditioned, then (4.1) is well-conditioned, and if (4.1) is ill-conditioned, then (4.2) is ill-conditioned. The following example shows that $\hat{K}_\infty(M)$ can be much larger than $K_\infty(M)$.

Example 4.1. Let $M = aI (a \geq 1)$. Straightforward calculation gives

$$\begin{aligned} \hat{K}_\infty(M) &\geq \left\| \begin{pmatrix} aI & -I \\ I & 0 \end{pmatrix} \right\|_\infty \left\| \begin{pmatrix} aI & -I \\ I & 0 \end{pmatrix}^{-1} \right\|_\infty \\ &= (1+a) \left\| \begin{pmatrix} 0 & I \\ -I & aI \end{pmatrix} \right\|_\infty = (1+a)^2 \end{aligned}$$

and

$$\begin{aligned} K_\infty(M) &= \max_{d\in[0,1]^n} \|(I-D+aD)^{-1}\|_\infty \|I-D+aD\|_\infty \\ &\leq \frac{\max_{0\leq\xi\leq 1} |(1+a\xi-\xi)|}{\min_{0\leq\xi\leq 1} |(1+a\xi-\xi)|} = a. \end{aligned}$$

For large a, $\hat{K}_\infty(M) - K_\infty(M)$ $(\geq a^2+a+1)$ is very large.

From Proposition 4.1 and Example 4.1, we may suggest that Newton-type methods for solving the nonlinear equations (1.8) have less perturbation error than Newton-type methods for (1.9). Now, we focus on Newton-type methods for (1.8). Obviously, it holds that

$$K(M) \geq \kappa(M)$$

as $e \in [0,1]^n$. For M being an H-matrix with positive diagonals, by Theorems 2.1 and 2.3 in [2], we have

$$K_\infty(M) \leq \max(1, \|M\|_\infty)\|\tilde{M}^{-1} \max(\Lambda, I)\|_\infty, \tag{4.5}$$

where Λ is the diagonal parts of M.

For M being an M-matrix with $\|M\|_\infty \geq 1$, we have

$$\kappa_\infty(M) \leq K_\infty(M) \leq \kappa_\infty(M)\|\max(\Lambda, I)\|_\infty. \tag{4.6}$$

Hence, the condition number $\kappa_\infty(M)$ is a measure of sensitivity of the solution of the system of linear equations for the worst case. Note that we have shown that $\kappa_\infty(M)$ is a measure of sensitivity of the solution of the LCP. Hence we may suggest that if Λ is not large, then the LCP is well-conditioned if and only if the system of linear equations (4.1) at each step of the Newton method is well-conditioned. Furthermore, for an M matrix, its diagonal elements are positive, and LCP$(\Lambda^{-1}M, \Lambda^{-1}q)$ and LCP(M, q) are equivalent. The inequalities in (4.6) yield $K_\infty(\Lambda^{-1}M) = \kappa_\infty(\Lambda^{-1}M)$.

5. Final remark. In [2], we provided the following error bound for the P-matrix LCP:

$$\|x - x(M, q)\| \leq \max_{d\in[0,1]^n} \|(I - D + DM)^{-1}\|\|r(x)\| \quad \text{for any } x \in R^n, \tag{5.1}$$

and we proved that (5.1) is sharper than the Mathias–Pang error bound [12]

$$\|x - x(M, q)\|_\infty \leq \frac{1 + \|M\|_\infty}{c(M)}\|r(x)\|_\infty \quad \text{for any } x \in R^n$$

in $\|\cdot\|_\infty$. Moreover, we showed that the error bound (5.1) can be computed easily for some special matrix LCP. For instance, if M is an H-matrix with positive diagonals, we have

$$\mu(M) := \max_{d\in[0,1]^n} \|(I - D + DM)^{-1}\| \leq \|\tilde{M}^{-1}\max(\Lambda, I)\|,$$

where Λ is the diagonal parts of M.

In this paper, we study the behavior of the solution $x(M, q)$ when there are some perturbations $\triangle M$ and $\triangle q$ in M and q. In particular, we show

$$\|x(M + \triangle M, q + \triangle q) - x(M, q)\| \leq \beta(M)\|\triangle M x(M + \triangle M, q + \triangle q) + \triangle q\|.$$

The constants $\mu(M)$ and $\beta(M)$ play different roles, where the former is for computation of error bounds and the latter is for sensitivity analysis.

Theorem 2.2 proves that $\beta(M)$ is smaller than the Mathias–Pang constant $1/c(M)$ for sensitivity and stability analysis [3]. Theorems 2.5 and 2.7 provide various interesting properties (1.5)–(1.7) of $\beta(M)$ when M is an H-matrix with positive diagonals, M-matrix, or symmetric positive definite matrix. These results show that the condition number $\kappa(M)$ is a measure of the sensitivity of the LCP(M, q). This means that if $\kappa(M)$ is small (large), then small changes in M or q result in small (large) changes in the solution $x(M, q)$ of the LCP(M, q).

When LCP(M, q) is used in the modeling of a practical application, the matrix M and vector q often contain errors due to inaccurate data, uncertain factors, etc. Hence, to make $x(M, q)$ useful and practical, it is very important to obtain some sensitivity information of the solution. This is one reason why sensitivity analysis of the LCP(M, q) has been studied so extensively [3]. On the web site http://www.st.hirosaki-u.ac.jp/∼chen/ExamplesLCP.pdf, we provide numerical examples including free boundary problems [14] and traffic equilibrium problems [3, 6] to illustrate the practical value of the new perturbation bounds (1.1)–(1.7).

Acknowledgments. The authors are grateful to Prof. Z. Q. Luo and two anonymous referees for their helpful comments.

REFERENCES

[1] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Classics Appl. Math. 9, SIAM, Philadelphia, 1994.

[2] X. Chen and S. Xiang, *Computation of error bounds for P-matrix linear complementarity problems*, Math. Program., 106 (2006), pp. 513–525.

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

[4] J. E. Dennis, Jr., and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Classics Appl. Math. 16, SIAM, Philadelphia, 1996.

[5] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, I *and* II, Springer-Verlag, New York, 2003.

[6] M. C. Ferris and J. S. Pang, *Engineering and economic applications of complementarity problems*, SIAM Rev., 39 (1997), pp. 669–713.

[7] S. A. Gabriel and J. J. Moré, *Smoothing of mixed complementarity problems*, in Complementarity and Variational Problems: State of the Art, M. C. Ferris and J.-S. Pang, eds., SIAM, Philadelphia, 1997, pp. 105–116.

[8] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, 1996.

[9] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.

[10] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.

[11] Z. Q. Luo, J.-S. Pang, and D. Ralph, *Mathematical Programs with Equilibrium Constraints*, Cambridge University Press, Cambridge, UK, 1996.

[12] R. Mathias and J.-S. Pang, *Error bounds for the linear complementarity problem with a P-matrix*, Linear Algebra Appl., 132 (1990), pp. 123–136.

[13] L. Qi, *Convergence analysis of some algorithms for solving nonsmooth equations*, Math. Oper. Res., 18 (1993), pp. 227–244.

[14] U. Schäfer, *An enclosure method for free boundary problems based on a linear complementarity problem with interval data*, Numer. Funct. Anal. Optim., 22 (2001), pp. 991–1011.

[15] U. Schäfer, *A linear complementarity problem with a P-matrix*, SIAM Rev., 46 (2004), pp. 189–201.

SIAM J. OPTIM.
Vol. 18, No. 4, pp. 1266–1285

STATISTICAL QUASI-NEWTON: A NEW LOOK AT LEAST CHANGE*

CHUANHAI LIU† AND SCOTT A. VANDER WIEL‡

Abstract. A new method for quasi-Newton minimization outperforms BFGS by combining least-change updates of the Hessian with step sizes estimated from a Wishart model of uncertainty. The Hessian update is in the Broyden family but uses a negative parameter, outside the convex range, that is usually regarded as the safe zone for Broyden updates. Although full Newton steps based on this update tend to be too long, excellent performance is obtained with shorter steps estimated from the Wishart model. In numerical comparisons to BFGS the new *statistical quasi-Newton (SQN)* algorithm typically converges with about 25% fewer iterations, functions, and gradient evaluations on the top 1/3 hardest unconstrained problems in the CUTE library. Typical improvement on the 1/3 easiest problems is about 5%. The framework used to derive SQN provides a simple way to understand differences among various Broyden updates such as BFGS and DFP and shows that these methods do not preserve accuracy of the Hessian, in a certain sense, while the new method does. In fact, BFGS, DFP, and all other updates with nonnegative Broyden parameters tend to inflate Hessian estimates, and this accounts for their observed propensity to correct eigenvalues that are too small more readily than eigenvalues that are too large. Numerical results on three new test functions validate these conclusions.

Key words. BFGS, DFP, negative Broyden family, Wishart model

AMS subject classifications. 65K10, 90C53

DOI. 10.1137/040614700

1. Introduction.

Quasi-Newton methods for unconstrained optimization are important computational tools in many scientific fields and are a standard subject in textbooks on computation. The BFGS method, proposed individually in [6], [14], [20], and [30], is implemented in most optimization software and is widely recognized as efficient. Generalizations of BFGS are available for large problems with memory limitations, for problems with bound constraints, and for a parallel computing environment. In theoretical investigations BFGS is known as a special case of the Broyden class [5]. Some Broyden updates with negative Broyden parameters have been found to produce faster convergence than BFGS updates [31], [8] but, for various reasons, have not been widely adopted. Indeed, Byrd et. al. conclude that "practical algorithms that preserve the excellent properties of the BFGS method are difficult to design." Nocedal and Wright [29] state that "the BFGS formula... is presently considered to be the most effective of all quasi-Newton updating formulae." In our opinion, BFGS remains the most popular front-runner because of two important unanswered

*Received by the editors September 9, 2004; accepted for publication (in revised form) April 9, 2007; published electronically October 24, 2007. Most of this work was done while the authors were in the Statistics Research Department, Bell Labs, Lucent Technologies. This work was performed by an employee of the U.S. Government or under U.S. Government contract. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.

http://www.siam.org/journals/siopt/18-4/61470.html

†Statistics and Data Mining Research, Bell Labs, 700 Mountain Avenue, Murray Hill, NJ 07974, and Department of Statistics, Purdue University, 150 N. University Street, West Lafayette, IN 47907 (chuanhai@stat.purdue.edu).

‡Statistics and Data Mining Research, Bell Labs, 700 Mountain Avenue, Murray Hill, NJ 07974, and Statistical Sciences Group, MS F600, Los Alamos National Laboratory, Los Alamos, NM 87545 (scottv@lanl.gov).

questions: What is the "best" negative Broyden parameter? and What initial step sizes should be used with negative Broyden parameters? This paper answers these questions by solving a least-change problem to approximate Newton directions and by estimating step sizes through a statistical model of Hessian uncertainty. We call the new algorithm *statistical quasi-Newton* (SQN).

1.1. Quasi-Newton methods. Quasi-Newton methods solve the unconstrained optimization problem

$$\min_x f(x), \qquad x \in \mathcal{R}^n,$$

in which both the objective function $f(x)$ and its gradient $g(x) \equiv \nabla f(x)$ are easy to compute but Newton's method is not applicable because direct evaluation of the Hessian matrix $G(x) \equiv \nabla^2 f(x)$ is practically infeasible. Quasi-Newton methods build up an approximate Hessian matrix using successive gradient evaluations. The general method iterates between a minimization (M-) step consisting of a one-dimensional search for a good point along an approximate Newton direction and an estimation (E-) step consisting of an update to the Hessian estimate. A more specific definition follows.

Generic quasi-Newton algorithm. Select a starting point $x_0 \in \mathcal{R}^n$ and a symmetric positive definite estimate, B_0, of the Hessian matrix $G(x_0)$. Evaluate $g_0 = g(x_0)$ and iterate over $k = 0, 1, 2, \ldots$ the following two steps.

M-Step. Search in the direction $-B_k^{-1} g_k$ for a step size $s_k > 0$ to obtain a new evaluation point and gradient,

$$x_{k+1} = x_k - s_k B_k^{-1} g_k, \qquad g_{k+1} = g(x_{k+1}),$$

that satisfy *the Wolfe conditions* for sufficient decrease of the function and for curvature (see (2) and (3) below).

E-Step. Estimate the Hessian matrix at x_{k+1} using the quantities B_k, x_k, x_{k+1}, g_k, and g_{k+1}. The estimate, B_{k+1}, must be symmetric and positive definite and must satisfy the *quasi-Newton condition*

$$B_{k+1}\delta_k = \gamma_k, \tag{1}$$

where

$$\delta_k \equiv x_{k+1} - x_k \quad \text{and} \quad \gamma_k \equiv g_{k+1} - g_k.$$

Condition (1) requires the vector of estimated second derivatives in the current step direction, $B_{k+1}\delta_k/s_k$, to agree with the corresponding numerical second derivatives γ_k/s_k. Various principles have been used to derive Hessian update formulae, but the general goal has been to minimize the change from B_k to B_{k+1} in some sense. This paper derives an update that minimizes change in a canonical sense and provides a model-based estimate for the step size s_k.

The Wolfe conditions referenced in the M-step are two standard requirements to ensure that sufficient progress is made toward the optimum even when the line search is not required to find the exact minimum in the given search direction. The Wolfe *sufficient decrease condition*,

$$f(x_{k+1}) \le f(x_k) - \rho_1 s_k g_k' B_k^{-1} g_k \qquad (\rho_1 \in (0,1), \text{ say } \rho_1 = 10^{-4}), \tag{2}$$

requires a reduction in $f(x)$ that is at least a fraction ρ_1 of that predicted by the directional derivative $-g_k B_k^{-1} g_k$. The Wolfe *strong curvature condition*,

$$|g'_{k+1}(B_k^{-1}g_k)| \le \rho_2 g'_k(B_k^{-1}g_k) \qquad (\rho_2 \in (\rho_1, 1), \text{ say } \rho_2 = 0.9), \tag{3}$$

requires at least a proportional decrease in the magnitude of the derivative in the search direction. Some algorithms impose a weaker curvature condition in which the absolute value is removed from the left-hand side of (3). Nocedal and Wright [29] discuss the importance of the Wolfe conditions in ensuring that sufficient progress is made on each iteration.

The best-known class of Hessian estimates used in the E-step are the rank-two Broyden updates [5]:

$$B_{k+1} = B_k - \frac{B_k \delta_k \delta'_k B_k}{\delta'_k B_k \delta_k} + \frac{\gamma_k \gamma'_k}{\delta'_k \gamma_k} + c_k \omega_k \omega'_k, \tag{4}$$

where

$$\omega_k \equiv \frac{\gamma_k}{\delta'_k \gamma_k} - \frac{B_k \delta_k}{\delta'_k B_k \delta_k} \tag{5}$$

and c_k is a scalar parameter to be specified. The usual parameterization takes $c_k = \phi_k\,(\delta'_k B_k \delta_k)$, where ϕ_k is known as the *Broyden parameter*. However, our exposition is more natural with the parameterization

$$c_k = (\lambda_k - 1)\,(\delta'_k \gamma_k)\,, \tag{6}$$

where the parameter λ_k is shown in section 3 to regulate the inflation of B_{k+1} relative to B_k. BFGS is the Broyden update with $\lambda_k = 1$ (i.e., $\phi_k = c_k = 0$).

There is a critical value λ_k^c such that B_{k+1} is positive definite for any $\lambda_k > \lambda_k^c \equiv 1 - r_k^{-1}$, where

$$r_k \equiv \frac{\gamma'_k B_k^{-1} \gamma_k}{\gamma'_k \delta_k} - \frac{\delta'_k \gamma_k}{\delta'_k B_k \delta_k}. \tag{7}$$

It can be shown that $r_k \ge 0$ by making use of the curvature condition (3) and the Cauchy–Schwarz inequality. If $r_k = 0$, then λ_k^c is taken to be $-\infty$.

1.2. Preview of SQN. The SQN method is remarkably simple and effective. This section briefly defines SQN and demonstrates its superiority to BFGS. Derivations and additional experimental results are provided in the following sections.

SQN algorithm. Follow the generic quasi-Newton algorithm with the following additional specifications. Initialize $\hat{s}_0 = 1$.
M-Step. Begin the line search from an initial evaluation point $x_k - \hat{s}_k B_k^{-1} g_k$.
E-Step. Estimate the Hessian using a Broyden update (4)–(6) with parameter

$$\lambda_k = \max\{0, 1 - (1-\epsilon) r_k^{-1}\}, \tag{8}$$

where ϵ is a small positive constant (e.g., $\epsilon = 10^{-6}$) and if $r_k = 0$, the max is taken to be 0. Estimate the next step size as

$$\hat{s}_{k+1} = \frac{g'_{k+1} B_{k+1}^{-1} g_{k+1}}{g'_{k+1} B_{k+1}^{-1} g_{k+1} + (1-\lambda_k)(\delta'_k \gamma_k)(g'_{k+1} B_{k+1}^{-1} \omega_k)^2} \;<\; 1. \tag{9}$$

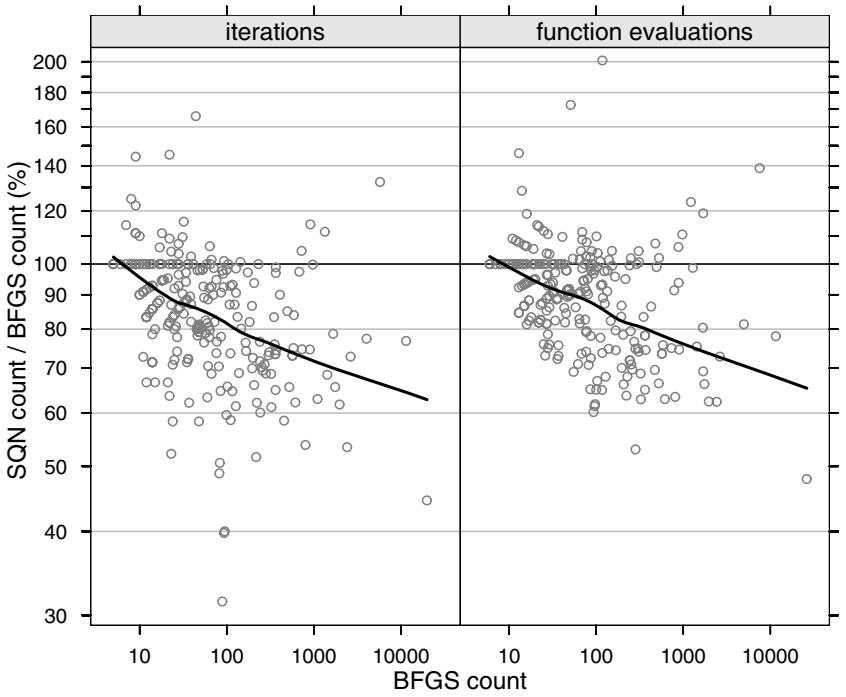


FIG. 1. *Improvement in SQN efficiency with problem difficulty for iterations, function evaluations, and gradient evaluations. Each point represents performance of SQN and BFGS on a given problem from the standard starting point.*

All the quantities needed to calculate $\hat{s}_{k+1}$ are readily available from the preceding M-step with no extra function or gradient evaluations required. Using $\epsilon > 0$ guarantees that B_{k+1} remains positive definite. The Broyden parameter corresponding to λ_k is $\phi_k = (\lambda_k - 1)\,(\delta_k'\gamma_k) \,/\, (\delta_k' B_k \delta_k)$, and this is negative because (3) implies $\delta_k'\gamma_k > 0$. Equation (9) is written in terms of the inverse Hessian estimate because Broyden updates are typically implemented on the inverse scale using the well-known dual form of (4). Similarly, $B_k\delta_k = -s_k g_k$ can be substituted into (7) for computational efficiency. See, for example, [29].

The shortened initial step size, $\hat{s}_k$, is crucial to improving the performance of Broyden updates with negative Broyden parameters. Zhang and Tewarson [31] use $\hat{s} = 1$ and comment that their negative Broyden algorithm improves iteration counts but that "less or no savings are achieved on the number of function evaluations" because initial steps are often too long to provide a sufficient decrease in the function value. SQN corrects this problem by effectively estimating the optimal step size for the given search direction.

Figure 1 shows that SQN typically converges with substantially fewer iterations and function evaluations than BFGS on 248 unconstrained optimization problems in the CUTE [3] suite. The left panel plots SQN iterations as a percent of BFGS iterations against BFGS iterations. The right panel shows the same information for function evaluations. SQN becomes more efficient relative to BFGS on the more difficult problems, as additional iterations offer additional opportunities for improvement. Performance on easy problems with few iterations is often dominated by the first iteration in which a poor choice of B_0 produces a poor search vector for any quasi-Newton

TABLE 1
Median percent improvement of SQN relative to BFGS by difficulty of problem.

	Easy	Medium	Hard
Iterations	8	14	26
Function evaluations	4	9	23

algorithm. In harder problems these start-up effects wash out so that the advantage of SQN over BFGS becomes more apparent. The trend curves in Figure 1 highlight this tendency. The smooth curves are robust local regressions [11] that follow the data without being unduly influenced by the low outlying points that would tend to make the trends even stronger.

Table 1 summarizes the improvement by splitting the test problems into three equal groups, *easy*, *medium*, and *hard*, according to the number of iterations for BFGS to converge. SQN's median improvement over BFGS is largest for the hardest 1/3 of the test problems, 25% in round numbers.

Our setup uses the line search [28] available from Argonne National Lab at ftp://info.mcs.anl.gov/pub/MINPACK-2/csrch in MINPACK-2. This line search evaluates the function value and gradient an equal number of times. The starting point x_0 is as given in the CUTE collection, and the initial Hessian estimate is $B_0 = c \cdot I_n$, where c is the geometric mean of the positive diagonal elements of the true Hessian at x_0. This is similar to the usual choice of $B_0 = I_n$, but scaling by c provides a more fair comparison because the true Hessian at x_0 tends to be much larger than I_n on the CUTE problems, and this gives an unfair advantage to BFGS, which has a bias toward inflating the Hessian estimate, as explained in section 3 below. For each test problem and starting point both SQN and BFGS are run until no valid step is found due to finite numerical precision. Then the best point x_* achieved by either algorithm is identified, and convergence is retrospectively declared at the first k for which

(10)
$$[f(x_k) - f(x_*)] + |(x_k - x_*)'g(x_*)| + |(x_k - x_*)'G(x_*)(x_k - x_*)| < 10^{-9}\,[1 + |f(x_*)|]\,.$$

This generalization of the assessment criterion [19] ensures that both the optima and the optimizers match.

The comparisons reported in Figure 1 and Table 1 are based on 248 problems in the CUTE collection. The test set consists of all unconstrained problems with maximum dimension of 500 that have continuous analytic second derivatives and compile with the default "large" version of the CUTE software. Of the 306 that fit these criteria, 15 appear to start at the optimum, 23 converge to a better local minimum with SQN than with BFGS, and 20 converge to a better minimum with BFGS. Removing these 58 cases leaves 248 test problems that support clean comparisons between SQN and BFGS.

We also conducted an initial study of the SQN algorithm, patterned after [31] using 20 of the test problems [27], each with 10 starting points. The results were similar: about 20% fewer iterations and gradient evaluations and about 10% fewer function evaluations compared to BFGS. This initial study used Fletcher's line search algorithm [15] with the tunable parameters set as suggested and utilizing his "sensible" choices for trial step lengths based on minimizing interpolating polynomials.

SQN compares favorably to other studies that have used negative Broyden parameters. Zhang and Tewarson [31] report 21% and 13% fewer iterations for their SDQN method relative to BFGS on problems of small and "increasing" dimension, respectively. However, their improvements were smaller using the EFE metric that

incorporates the number of function evaluations. Byrd, Liu, and Nocedal [8] report improvements of 18% on iterations and 12% on function evaluations for a smaller set of tests using their Method I, which is not practical as a quasi-Newton update because it requires evaluation of $G(x)$.

The remainder of the article is arranged as follows. Section 2 gives a select history of ideas in quasi-Newton development with emphasis on the least-change principle and argues for a particular scale-free matrix as the most appropriate measure of the change between consecutive Hessian estimates. Section 3 introduces a transformation into canonical coordinates, derives (4)–(8) as the new least-change update, and shows that it preserves Hessian accuracy from one iteration to the next in a certain sense. Section 4 introduces a Wishart model to describe Hessian uncertainty and derives (9) as an estimate of the optimal step size. Section 5 compares performance on three new test functions designed to verify our understanding of why SQN is better than other Broyden methods. Section 6 explores connections to other least-change derivations and mentions ideas for future research.

2. Least-change updates. Fletcher's overview [17] of methods for unconstrained optimization is an excellent introduction to the huge literature on quasi-Newton methods. This section briefly reviews the historical ideas that led to the least-change principle on which the most influential quasi-Newton methods are based. A line of reasoning is then given to suggest a certain relative-change matrix as being the most appropriate measure of change for the goal of approximating Newton search directions. This leads to the SQN update that was introduced in section 1.2. Although the SQN update happens to be in the Broyden class, it is derived in section 3 by minimizing change over *all possible* quasi-Newton updates.

2.1. Historical developments. Crockett and Chernoff [12] stated the idea of building up a Hessian estimate iteratively so as to approximate the Newton method:

> *..., it is possible to obtain, from the successive approximations, certain relevant information about terms of order higher than those actually computed, and to conveniently use this information to improve the rate of convergence.*

The basic idea of Broyden [4] as articulated in [7] was that the Hessian update *"should therefore require, if possible,...,* no change *to B_k in any direction orthogonal to δ_k."* Broyden was solving a system of differential equations, and his mathematical formulation [$B_{k+1}\delta_k = \gamma_k$ and $(B_{k+1} - B_k)q = 0 \;\; \forall q : q'\delta_k = 0$] produces an asymmetric update that is not appropriate for the problem $\min f(x)$.

Taking a more mathematical approach, Broyden [5] dropped the "orthogonality" part of his original intuition and sought instead a low-rank Hessian update. This led to the Broyden class (4) of symmetric rank-two updates. Subsequent researchers also focused on making small modifications to the Hessian without explicit concern for the space orthogonal to the search direction. Greenstadt [21], for example, wrote,

> *Let us ask for the "best" correction in some sense. There are many possible choices to make, but a good one is to ask for the* smallest *correction, in the sense of some norm. To a certain extent, this would tend to keep the elements of $[B_k^{-1}]$ from growing too large, which might cause an undesirable instability.*

The extensive review [25] emphasizes the importance of the *least-change principle* in deriving many of the most effective quasi-Newton methods.

The important special case of a Broyden update with $\lambda_k = 1$ is called BFGS after the four authors who individually published the update formula in 1970. Goldfarb [20]

worked with the scaled difference of inverse Hessian estimates

$$E_W^* \equiv W^{1/2}\left(B_{k+1}^{-1} - B_k^{-1}\right)W^{1/2}, \tag{11}$$

where the symmetric matrix W satisfies $W\delta_k = \gamma_k$. He derived the BFGS update by using the results in [21] to minimize the Frobenius norm $\|E_W^*\|_F \equiv [\mathrm{tr}\,(E_W^* E_W^*)]^{1/2}$ over the class of symmetric matrices B_{k+1} that satisfy the Newton condition (1). Thus, BFGS is a least-change update. But the metric of change is important. For example, using the same W but minimizing the Frobenius norm of

$$E_W \equiv W^{-1/2}\left(B_{k+1} - B_k\right)W^{-1/2} \tag{12}$$

produces the Broyden update with $\lambda_k = 1+\delta_k' B_k \delta_k/(\delta_k'\gamma_k)$. This is known as DFP [13], [18] and is generally regarded as inferior to BFGS.

Fletcher [14] advocated restricting attention to Broyden updates that are convex combinations of the BFGS and DFP updates because such updates satisfy a monotone eigenvalue property when used to minimize quadratic functions. Recently, however, various choices of negative Broyden parameters ($\phi_k < 0$ corresponding to $\lambda_k < 1$) have been studied. See, for example, [31], [8], [23], [17], and [26]. These authors report that negative Broyden parameters can reduce iteration counts, although in some cases this comes at the cost of increased numbers of function evaluations. The potential for improvement relative to BFGS seems to be best if the initial Hessian estimate is much too large. Robust improvement over BFGS has been elusive. Indeed, Zhang and Tewarson [31] concluded that such investigations have not shaken the position of BFGS as the most popular front-runner.

2.2. A new measure for least change. Minimizing the change from B_k to B_{k+1} is a generally accepted principle. There is no agreement, however, on how to measure that change. Zhao [32] derives 10 different optimal updates by considering five possible matrix norms applied to two different matrices that measure change. The function for measuring change is empirically important: BFGS outperforms DFP even though the two are least-change duals derived from E_W^* and E_W, respectively.

A sensible matrix measure of change that has received little attention in the literature is the difference $B_{k+1} - B_k$ scaled by the current estimate B_k, namely

$$E_B \equiv B_k^{-1/2}\left(B_{k+1} - B_k\right)B_k^{-1/2}. \tag{13}$$

Normalizing a difference is appropriate because, in every direction, E_B measures change of the Hessian estimate *relative to current nominal value*, and this produces a scale-free method. Greenstadt [22] states that such normalization "renders harmless the accidents of coordinate selection in a given problem." One possible danger in making E_B small is that B_{k+1} could become singular (or even indefinite if allowed), and this could produce unstable quasi-Newton search vectors, based on B_{k+1}^{-1}. However, applying no direct penalty to large differences on the inverse scale is more aggressive than BFGS, in the same spirit as employing negative values of the Broyden parameter. In fact, the next section will demonstrate that minimizing $\|E_B\|_F$ produces exactly a negative Broyden update.

Interestingly, minimizing E_B (with respect to commonly used scalar measures of matrices) is equivalent to minimizing

$$E_B^* \equiv B_{k+1}^{1/2}(B_k^{-1} - B_{k+1}^{-1})B_{k+1}^{1/2} \tag{14}$$

because E_B^* and E_B have the same eigenvalues, as shown in Appendix C. The matrix E_B^* scales the difference in inverse estimates by the still-to-be-determined update. Greenstadt [21] minimized a weighted change of the inverse estimates. In deriving BFGS, Goldfarb [20] writes, "If, instead, $[B_{k+1}]$ is substituted for [the weight matrix] in [Greenstadt's result], then [BFGS] is obtained." Although this sounds like minimizing E_B^*, Goldfarb in fact minimized E_W^* with a fixed weight matrix that satisfied the quasi-Newton condition required of B_{k+1}, namely $W\delta_k = \gamma_k$. SQN, on the other hand, can be viewed as directly using the unknown B_{k+1} as the weight matrix in Greenstadt's objective function.

3. SQN: Least relative change. The form of E_B in (13) as a measure of change motivates transforming the coordinates of x by $B_k^{1/2}$ so that the problem of updating the Hessian estimate takes a simple form. This section uses Broyden's original idea of making no change to the portion of B_k that is orthogonal to δ_k but applies the idea in a transformed coordinate system.

As the focus is on the kth step of the quasi-Newton algorithm, the notation is streamlined from this point forward by dropping subscripts k and replacing subscripts $k+1$ by "+."

3.1. Canonical coordinates. For conceptual convenience, at the kth iteration transform x in such a way that the line search is along the first component direction and the current Hessian estimate B transforms to the identity matrix. This is accomplished by the linear transformation

$$\tilde{x} = U'B^{1/2}x, \tag{15}$$

where U is an orthonormal rotation matrix with the first column equal to $B^{1/2}\delta(\delta'B\delta)^{-1/2}$. In the transformed space the current step is strictly along the first component direction:

$$\tilde{x}_+ - \tilde{x} = (\delta'B\delta)^{1/2}(1, 0, \ldots, 0)'.$$

The objective function and gradient become

$$\tilde{f}(\tilde{x}) \equiv f(x) \quad \text{and} \quad \tilde{g}(\tilde{x}) \equiv \nabla\tilde{f}(\tilde{x}) = U'B^{-1/2}g(x),$$

and the transformed Hessian is

$$\tilde{G}(\tilde{x}) \equiv \nabla^2\tilde{f}(\tilde{x}) = U'B^{-1/2}G(x)B^{-1/2}U. \tag{16}$$

Substituting the estimated Hessian B for $G(x)$ in (16) produces the transformed estimate $\tilde{B} = I_n$, the n-dimensional identity matrix.

3.2. Observed and missing information. Define second-order numerical derivatives of $\tilde{f}(\tilde{x})$ along the search direction as

$$\begin{bmatrix} a \\ b \end{bmatrix} \equiv \frac{\tilde{g}(\tilde{x}_+) - \tilde{g}(\tilde{x})}{(1, 0, \ldots, 0)(\tilde{x}_+ - \tilde{x})} = \frac{U'B^{-1/2}\gamma}{(\delta'B\delta)^{1/2}}, \tag{17}$$

where the first element a is a scalar and b is an $(n-1)$-dimensional vector. The curvature condition (3) implies that $a \geq (1-\rho_2)/s > 0$. The quasi-Newton condition (1) is equivalent to the intuitive idea that the numerical derivatives (17) form the first

column of the updated Hessian matrix. Since the Hessian is symmetric, the general form of update in transformed coordinates becomes

$$\tilde{B}_+ = \begin{bmatrix} a & b' \\ b & C \end{bmatrix}, \tag{18}$$

where symmetric C is to be determined subject only to the constraint $\tilde{B}_+ > 0$, which is equivalent to $C - a^{-1}bb' > 0$. (The notation $M > 0$ indicates that the matrix M is positive definite.) C represents curvature in the complimentary space, that is, the space canonically orthogonal to the current search direction.

Following Broyden's idea that no information is gained in directions orthogonal to δ suggests the updating scheme obtained by taking $C = I_{n-1}$ if doing so produces $\tilde{B}_+ > 0$, i.e., if $a > b'b$. But, what does one do if $a \leq b'b$? The question itself implies that certain information on C *is* provided by the observed data (a, b) along with the assumption that the Hessian matrix is positive definite. In general, C should be a function of a and b.

The following theorem provides the least-change update based on the Frobenius norm of E_B.

THEOREM 1 (SQN update). *The quasi-Newton update that minimizes* $\|E_B\|_F$ *(and hence also* $\|E_B^*\|_F$*) subject to* (1) *and* $B_+ \geq 0$ *has canonical form*

$$\tilde{B}_+ = \begin{bmatrix} a & b' \\ b & I_{n-1} + \lambda_{\text{SQN}}\, bb'/a \end{bmatrix}, \tag{19}$$

where, for $\tilde{r} \equiv b'b/a$,

$$\lambda_{\text{SQN}} = \begin{cases} 0 & \text{if } \tilde{r} \leq 1, \\ 1 - \tilde{r}^{-1} & \text{otherwise.} \end{cases} \tag{20}$$

$\tilde{B}_+$ *is singular for* $\tilde{r} \geq 1$.

Proof. Appendix C implies that $\|E_B\|_F = \|E_B^*\|_F$:

$$\begin{aligned} \|E_B\|_F^2 &= \left\| B^{-1/2}\left(B_+ - B\right)B^{-1/2} \right\|_F^2 = \left\| U'B^{-1/2}B_+B^{-1/2}U - I_n \right\|_F^2 = \left\| \tilde{B}_+ - I_n \right\|_F^2 \\ &= \operatorname{tr}\left(\left[\begin{pmatrix} a & b' \\ b & C \end{pmatrix} - I \right] \left[\begin{pmatrix} a & b' \\ b & C \end{pmatrix} - I \right] \right) \\ &= \operatorname{tr}\left(\Phi^2\right) - 2\operatorname{tr}\left(\Phi\right) + 2b'\Phi b/a + (a + \tilde{r})(a + \tilde{r} - 2) + n, \end{aligned}$$

where $\Phi \equiv C - bb'/a$ and we have used $\tilde{B}_+ = U'B^{-1/2}B_+B^{-1/2}U$ from (16). $B_+ \geq 0$ is equivalent to $\Phi \geq 0$ so that minimizing $\|E_B\|_F$ over $B_+ \geq 0$ is equivalent to minimizing the first three terms of the final expression over $\Phi \geq 0$.

Denote the eigenvalues of Φ by $0 \leq \eta_1 \leq \cdots \leq \eta_n$. Then

$$\operatorname{tr}\left(\Phi^2\right) - 2\operatorname{tr}\left(\Phi\right) + 2b'\Phi b/a \;\geq\; \sum_i \eta_i^2 - 2\sum_i \eta_i + 2\eta_1 b'b/a \tag{21}$$

with equality if and only if the first eigenvector of Φ is proportional to b. The right-hand side of (21) is minimized by $\eta_2 = \cdots = \eta_n = 1$ and

$$\eta_1 = \max\left\{0, 1 - \tilde{r}\right\}.$$

Thus, the left-hand side of (21) is minimized by a matrix with the specified eigenvalues and first eigenvector equal to $b/\sqrt{b'b}$. The required matrix is

$$\Phi = I + \left[\frac{\max\{0, 1-\tilde{r}\} - 1}{\tilde{r}}\right] \frac{bb'}{a},$$

and this corresponds to the optimal C given in the theorem. □

Theorem 1 demonstrates that making no change in the complementary space (i.e., $C = I_{n-1}$) does, in fact, produce a least-change update. The theorem also provides a larger estimate for C when needed to preserve nonnegative definiteness. Our implementation of SQN uses a safeguarded choice of λ_{SQN} to prevent the Hessian estimate from becoming singular. See (8).

Behind the intuition that one should make small alterations to the Hessian estimate in the complementary space lies a principle that accuracy obtained on previous iterations should be preserved as much as possible. The following proposition demonstrates that the SQN update achieves the goal of preserving Hessian accuracy in a certain sense.

PROPOSITION 1 (SQN accuracy preservation). *If the true Hessian in canonical coordinates is positive definite and given by*

$$\tilde{B}_+ = \begin{bmatrix} a & b' \\ b & C_{\text{TRUE}} \end{bmatrix}, \tag{22}$$

then $C_{\text{SQN}} \equiv I_{n-1} + \lambda_{\text{SQN}} bb'/a$ is at least as accurate as I_{n-1} for estimating C_{TRUE} in any direction either parallel to b or orthogonal to b. That is,

$$\left|u'\left(C_{\text{SQN}} - C_{\text{TRUE}}\right)u\right| \;\le\; \left|u'\left(I_{n-1} - C_{\text{TRUE}}\right)u\right| \tag{23}$$

for any u such that either $u'b = 0$ or $u \propto b$. Furthermore, this is not necessarily true for any larger estimate $\hat{C} = C_{\text{SQN}} + VV'$, where V is any nonzero matrix with $n-1$ rows.

See Appendix A for a proof.

The following proposition provides the canonical form for the well-known Broyden family and shows that SQN updates are particular members.

PROPOSITION 2 (canonical Broyden updates). *Under the canonical transform* (15) *the Broyden update* (4) *transforms to*

$$\tilde{B}_+ = \begin{bmatrix} a & b' \\ b & I_{n-1} + \lambda bb'/a \end{bmatrix}, \tag{24}$$

where $\lambda = 1 + c/(\delta'\gamma)$. In particular, the usual Broyden parameter is $\phi = (\lambda - 1)a$, and important special cases are given as follows:

Method	λ	ϕ
SQN	$\max\{0, 1-\tilde{r}^{-1}\}$	$\max\{-a, -a\tilde{r}^{-1}\}$,
BFGS	1	0
DFP	$1 + a^{-1}$	1

where if $\tilde{r} = 0$, the max *is taken to be the first argument.*

See Appendix B for a proof. The proof also provides formulae for a and b in terms of the usual quantities δ, γ, and B and shows that $\tilde{r} = r$, where $\tilde{r} = b'b/a$ is defined in Theorem 1 and r is given in (7).

Although BFGS minimizes several different measures of change [16], Proposition 2 indicates that BFGS *increases* the lower right block ($\lambda > 0$) over its previous value of I_{n-1}, whereas SQN leaves it unchanged if possible, or adds a fraction of the BFGS correction in order to preserve positive semidefiniteness. The conclusion of Proposition 1 is that neither BFGS nor DFP preserves accuracy of the previous Hessian estimate (in the canonical sense of (23)) over a large class of directions. It is interesting that DFP explodes as a becomes small.

4. Step size estimation. In trial experiments with the SQN update, we carried out the quasi-Newton M-step using a line search in which the initial step size was unity; that is, the line search used an initial evaluation point of $x - \hat{s}B^{-1}g$ with $\hat{s} = 1$, which is the Newton step under the assumption that B is the actual Hessian. The experiments demonstrated that the SQN update tended to reduce the number of iterations to convergence compared to BFGS but did not consistently reduce the number of function evaluations required. Further investigation showed the reason: unit steps are often too long when the SQN update is used. The steepest descent method (SDQN) [31] also uses negative Broyden parameters, and they state, "*SDQN tends to give steps longer than BFGS steps, and therefore is more likely to violate the [sufficient decrease] condition.*" When unit steps are used, fewer iterations seem to come with the price of more function evaluations per iteration. Some numerical results with unit step sizes on SQN and other Broyden updates are reported in section 5.

Why do negative Broyden parameters produce steps that are too long? A rough explanation is that a negative Broyden parameter produces a smaller Hessian estimate than BFGS. Compare $\lambda < 1$ in Proposition 2 with $\lambda = 1$. A smaller B implies a longer unit step $-B^{-1}g$. Therefore, if unit steps are suitable for BFGS, then unit steps may well be too long for use with negative Broyden parameters. This reasoning is admittedly rough; it does not account for differences in the step *direction* and does not provide guidance for selecting more appropriate step sizes. This section proposes a Wishart model to describe uncertainty of the unknown Hessian and then derives an estimate of the optimal step size as a function of the Broyden parameter used in updating the Hessian. The SQN initial step size (9) is a special case.

4.1. A Wishart model for the Hessian matrix. The unknown Hessian $\tilde{G}_+ = \tilde{G}(\tilde{x}_+)$ can be modeled as a random matrix whose probability distribution quantifies the plausibility of all possible canonical Hessians. It is reasonable to use a probability model for this purpose because the true Hessian varies unpredictably from one quasi-Newton iteration to the next and from one objective function to another. Therefore $\tilde{G}_+$ is never completely known. Furthermore, modeling $\tilde{G}_+$ as a random matrix provides a means of incorporating new curvature information obtained in a line search and appropriately updating the distribution of the unknown Hessian. The updated distribution is the key to determining what length of step should be taken in any given direction.

Several properties are desirable for the distribution of $\tilde{G}_+$. It should

(i) be centered at the previous estimate $\tilde{B} = I_n$,

(ii) have probabilities that taper off toward zero for matrices far from I_n, and

(iii) describe equal uncertainty in every direction because, although $\tilde{G}_+$ is likely less uncertain in the directions of recent steps, these directions are not available for use within the quasi-Newton framework.

The simplest statistical model for symmetric positive definite matrices that has the above properties is the Wishart distribution with expectation I_n:

$$\nu\tilde{G}_+ \sim \text{Wishart}_n\left(I_n, \nu\right), \tag{25}$$

where $\nu \geq n+1$ is the degrees of freedom parameter. The distribution of $\tilde{G}_+$ becomes more concentrated around I_n as ν increases. See, e.g., [2] for the definition and properties of the Wishart family. The probability density function of $\tilde{G}_+$ is proportional to

$$\left|\tilde{G}_+\right|^{(\nu-n-1)/2} \exp\left\{-\frac{\nu}{2}\text{tr}\left(\tilde{G}_+\right)\right\}. \tag{26}$$

Because (26) involves only $\tilde{G}_+$ through its determinant and trace, any orthogonal rotation, $R'\tilde{G}_+R$ where $R'R = I_n$, is distributed identically to $\tilde{G}_+$. This *directional symmetry* seems an appropriate requirement for modeling the Hessian in canonical coordinates.

In the quasi-Newton framework, the first row and column of $\tilde{G}_+$ are considered to be known from the numerical second derivatives (17). Therefore

$$\tilde{G}_+ = \begin{bmatrix} a & b' \\ b & C \end{bmatrix}, \tag{27}$$

where a and b are observed and C is not. Standard Wishart theory (see, e.g., [2]) provides the conditional distribution $[C|a,b]$ through

$$\nu\left[C - \frac{bb'}{a}\middle| a, b\right] \sim \text{Wishart}_{n-1}\left(I_{n-1}, \nu-1\right).$$

The conditional expectation and mode are

$$\text{E}(C|a,b) = \frac{\nu-1}{\nu} I_{n-1} + \frac{bb'}{a}, \tag{28}$$

$$\text{Mode}(C|a,b) = \frac{\nu}{\nu-n-1} I_{n-1} + \frac{bb'}{a}. \tag{29}$$

The two multipliers on I_{n-1} depend on the degrees of freedom, ν, and they differ because the Wishart model is skewed toward large positive definite matrices. But both coefficients approach unity as $\nu \to \infty$, the *large-sample* limit. Although ν could be estimated from a and b, using the large-sample limit is an attractive simplification that corresponds to modeling the current Hessian estimate as arbitrarily accurate *before* observing a and b.

Comparing (28) and (29) to (24) in Proposition 2 shows that the large-sample conditional expectation and mode under a Wishart model are exactly equal to the BFGS update. Specifically, let $B_+(\lambda)$ denote the Broyden update (4)–(6) with parameter λ and let $\tilde{B}_+(\lambda)$ denote the corresponding canonical form given by (24). Then

$$\begin{aligned} \lim_{\nu\to\infty} E\left(G_+|a,b\right) &= B^{1/2}U\left[\lim_{\nu\to\infty} E\left(\tilde{G}_+\middle| a,b\right)\right]U'B^{1/2} \\ &= B^{1/2}U\tilde{B}_+(1)U'B^{1/2} \\ &= B_+(1), \end{aligned} \tag{30}$$

which is the BFGS update.

Although (30) is the simplest statistical estimate of the Hessian, the SQN update is a better choice for *sequential* Hessian estimation because it preserves accuracy obtained in previous iterations (section 2.2 and Proposition 1). When estimating the optimal step size, however, accuracy preservation is not a concern—an appropriate step size in one iteration may or may not be appropriate in the next. Thus, while the SQN Hessian update is derived to preserve accuracy, the SQN step size, derived in the next section, uses conditional expectation to estimate the optimal step, given the most recent curvature information.

4.2. Optimal step size. An estimate of the optimal step size for any given Broyden update can be derived from the Wishart model. Let d_+ represent an arbitrary search direction to be taken in the M-step on iteration $k+1$. A second-order Taylor expansion of $f(\cdot)$ about the point x_+ gives the quadratic approximation

$$f(x_+ + sd_+) \approx f(x_+) + sd'_+g_+ + \frac{s^2}{2}d'_+G_+d_+ \tag{31}$$

with optimum step size

$$s^* = \frac{-d'_+g_+}{d'_+G_+d_+} \tag{32}$$

obtained by differentiating (31) with respect to s and setting the result to zero. The denominator of (32) involves the unknown Hessian, but an estimate of s^* can be obtained by replacing G_+ with its large-sample conditional expectation from (30):

$$\lim_{\nu\to\infty} E\left(G_+|a,b\right) = B_+(1) = B_+(\lambda) + (1-\lambda)(\delta'\gamma)\omega\omega', \tag{33}$$

where (4) has been used to express $B_+(1)$ in terms of a general Broyden update. The resulting optimum step size is obtained by plugging (33) into (32) and taking $d_+ = -B_+^{-1}(\lambda)g_+$, the next quasi-Newton step direction:

$$\hat{s}(\lambda) = \frac{g'_+B_+^{-1}(\lambda)g_+}{g'_+B_+^{-1}(\lambda)g_+ + (1-\lambda)(\delta'\gamma)(g'_+B_+^{-1}(\lambda)\omega)^2}. \tag{34}$$

This is the step size formula (9) of the SQN algorithm. For BFGS ($\lambda = 1$), the estimated optimum is $\hat{s}(1) = 1$, which suggests that unit steps may work better for BFGS than for any other Broyden update.

Results comparing BFGS to the SQN algorithm using (34) are shown in Figure 1 and demonstrate that SQN achieves consistent reduction in function evaluations, as well as iteration counts, compared to BFGS. Additional comparisons to SQN with unit initial steps for three new test functions are reported next.

5. Results on three new test functions. The CUTE test problems have become standard for comparing quasi-Newton algorithms, but they are not particularly useful for empirically validating our claim that BFGS tends to inflate B_k and that SQN is more neutral. This section uses three new test functions for that purpose.

It was found that BFGS is lopsided [8]: it can more readily increase Hessian estimates that are too small than shrink ones that are too large. This was surprising in light of the strong "self-correcting" property of the BFGS update that was established [10]: the relative error between the curvature predicted by B_k and the curvature observed in the current line search is transmitted exactly to the relative change of the

TABLE 2
Three test functions $f(x) = \sum_1^4 f_i(x_i)$ with simple Hessians.

	$f_i(x_i)$	$G_{ii}(x)$	Anticipated best λ_{NOM}
f^{dec} :	$\frac{1}{2}x_i^2 + \frac{1}{12}\eta_i^2 x_i^4$	$1 + (\eta_i x_i)^2$	negative
f^{inc} :	$\eta_i^{-2}\left[\eta_i x_i \arctan(\eta_i x_i) - \frac{1}{2}\ln(1+\eta_i^2 x_i^2)\right]$	$[1 + (\eta_i x_i)^2]^{-1}$	positive
f^{sin} :	$\frac{1}{2}x_i^2 + \eta_i^{-2}\left[\eta_i x_i - \sin(\eta_i x_i)\right]$	$1 + \sin(\eta_i x_i)$	near zero

determinant from $|B_k|$ to $|B_{k+1}|$. Proposition 2, on the other hand, shows that BFGS corrections actually inflate B_k in the space canonically orthogonal to the search direction, whereas SQN corrections leave that part of the Hessian unchanged (subject to positive definiteness) and therefore should cope equally well with estimates that need to shrink as with ones that need to grow. Furthermore, choosing λ_k to be less than 0 or greater than 1 should make these effects more pronounced.

To check this understanding, we employ three new test functions f^{dec}, f^{inc}, and f^{sin} with simple Hessians that respectively decrease, increase, and change sinusoidally as x_k moves in the direction of the optimum value. Each function has $n = 4$ dimensions and has the form $f(x) = \sum_1^4 f_i(x_i)$. The functions are defined in Table 2, where the values $(\eta_1, \eta_2, \eta_3, \eta_4) \equiv (1, 2, 4, 8)$ scale how quickly curvature changes in each coordinate direction. Each function is convex and has a diagonal Hessian with an ith diagonal element as listed in the table. In each case the minimizer is $x^* = (0, 0, 0, 0)$, $f(x^*) = 0$, and $G(x^*) = I_4$.

For these functions we implement a range of Broyden updates with

$$\lambda = \max\left\{\lambda_{\text{NOM}}, 1 - (1-\epsilon)r^{-1}\right\},$$

where λ_{NOM} is set between -2 and 3, $\epsilon = 10^{-6}$, and initial step sizes are given by (34). Special cases are $\lambda_{\text{NOM}} = 0$ and 1, which correspond to SQN and BFGS, respectively.

The rationale for testing with functions whose Hessians change monotonically (f^{dec}, f^{inc}) or unpredictably (f^{sin}) is to verify our claim that BFGS needlessly inflates the previous Hessian estimate whereas SQN treats it neutrally. With f^{inc}, for example, the most appropriate Hessian estimate in iteration $k+1$ will tend to be larger than in iteration k. BFGS could have an advantage over SQN because it tends to inflate the Hessian beyond its previous value in the complementary space. In this case, the best choice of λ_{NOM} should be larger than 0 and possibly even larger than 1, the BFGS value. For f^{dec}, on the other hand, SQN should have the advantage over BFGS and the optimal λ_{NOM} should be negative. For f^{sin}, there is no consistent pattern for the Hessian on one step compared to the previous step so that $\lambda_{\text{NOM}} = 0$ (i.e., SQN) should be nearly optimal. In each case, more extreme values of λ_{NOM} should produce more extreme effects.

5.1. Results for different λ_{NOM}. Figure 2 plots average counts to convergence as a function of λ_{NOM} with each panel representing one of the new test functions. Each plotted symbol represents an average count over 1000 random starting points. The vertical scales are set to support relative comparisons, the most obvious of which is that λ_{NOM} has the greatest effect for f^{dec} and the least for f^{sin}. Iterations, function evaluations, and gradient evaluations are shown using different plotting symbols. Fletcher's line search [15], as discussed in section 1.2, was used in this study. The true value is used for the starting Hessian estimate, $B_0 = G(x_0)$.

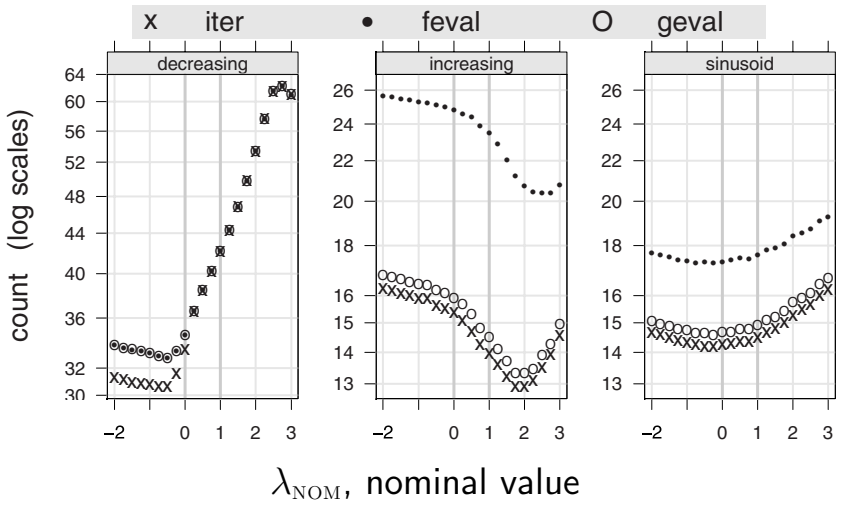


FIG. 2. *Performance counts versus* λ_{NOM} *on test functions with Hessians that are decreasing, increasing, and sinusoidal as* x_k *moves toward the minimum. Different symbols are used for iterations (*$\times$ *iter), function evaluations (*• *feval), and gradient evaluations (* O *geval). Initial steps are estimated using* (9). *The special value* $\lambda_{\text{NOM}} = 0$ *is SQN, and* $\lambda_{\text{NOM}} = 1$ *is BFGS.*

The starting points x_0 were chosen at random in such a way that they tend to be oriented in the direction of $(\eta_1, \eta_2, \eta_3, \eta_4)$. Specifically, the ith component of x_0 was drawn randomly as

$$x_{0,i} = K\eta_i(1 + z_i/3),$$

where the z_i are independent $\mathrm{N}(0,1)$ random variables and the scale was set as $K = 200$ for f^{dec}, $K = 50$ for f^{inc}, and $K = 1000\|\eta\|$ for f^{sin}. These choices reflect a little experimentation aimed at producing differences between BFGS and SQN that are large enough to be interesting without requiring unwieldy numbers of iterations. As far as we know, other choices produce similar results, though we have not studied this extensively. Convergence was declared when $f(x_k) < 10^{-10}$.

For f^{dec}, Figure 2 demonstrates that SQN is indeed better able to cope with a decreasing Hessian than BFGS, and further improvement is obtained by using slightly negative values of λ_{NOM}. The situation is reversed for f^{inc}. BFGS handles the increasing Hessian better than SQN, and further improvement is obtained by taking λ_{NOM} as large as 2. Finally, for f^{sin} the Hessian changes arbitrarily, and the SQN update ($\lambda_{\text{NOM}} = 0$) is nearly optimal.

Several additional comments on these results are worth noting. First, within each panel all three curves have nearly the same shape. But on f^{dec} function evaluations are always equal to gradient evaluations, whereas function evaluations are substantially higher on f^{inc} and f^{sin}. This indicates that the initial step size estimate is better for f^{dec} than for the other two functions because, with the Fletcher line search, if initial step sizes are too large to produce a sufficient decrease in the function value, then the function is reevaluated at additional trial steps with no gradient evaluations. Second, for any $\lambda_{\text{NOM}} < 1$ some values of λ_k will likely exceed λ_{NOM} because of the requirement that B_{k+1} remain positive definite. This produces an asymmetry in the results so that the performance differences between $\lambda_{\text{NOM}} = -1$ and 0 are not as great as the differences between 0 and 1. In fact, our selection of starting points that

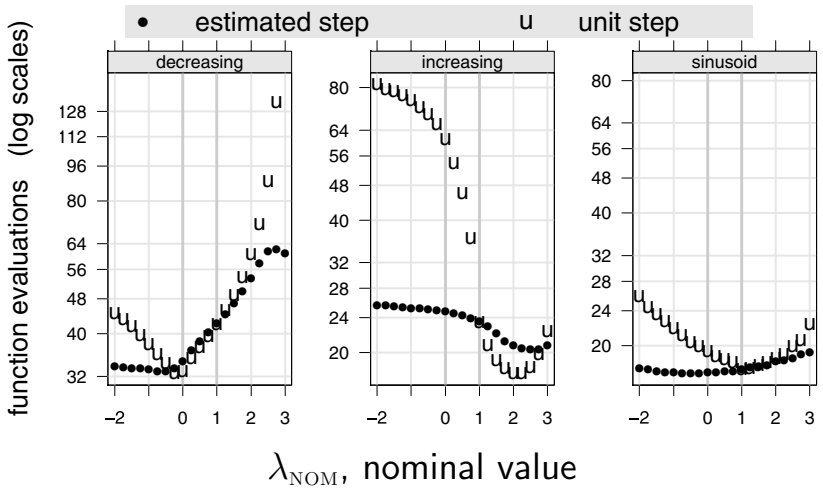


FIG. 3. *Function evaluation counts versus* λ_{NOM} *for three test functions. The plots compare performance with unit initial steps* (u) *against estimated initial steps* (•) *using* (9). *The dots in this figure are the same as in Figure* 2.

are biased in the direction of η was made to enhance the effect of λ_{NOM} below 1 on $f^{\mathrm{inc}}$and f^{sin}. The patterns in Figure 2 are smooth because they average across 1000 starting points. If counts from a single starting point were plotted, the patterns for f^{inc} and f^{sin} would be virtually impossible to discern because of noise in the data. Thus, it would be meaningless to compare different choices of λ_{NOM} based on only a few test cases.

5.2. Results for different step sizes. Figure 3 demonstrates the importance of using estimated step sizes, especially with $\lambda_{\mathrm{NOM}} < 1$. The experiment is the same as in Figure 2, except that the algorithm was also run with unit initial step sizes. The plots compare average function evaluation counts for unit initial steps against those for estimated steps. In each panel, as λ_{NOM} decreases from 1 (BFGS), the unit initial step results eventually become much worse than the results with estimated steps. The same appears to be true as λ_{NOM} becomes positive and large. The curves intersect at $\lambda_{\mathrm{NOM}} = 1$ because the estimated step size is 1.

At $\lambda_{\mathrm{NOM}} = 0$ (SQN) the results of Figure 3 are most revealing on f^{inc}. In this case the SQN Hessian estimate tends to be too small so that unit step sizes are too large. Estimated step sizes are smaller and perform much better, although they may still be too large, as indicated in Figure 2, by the gap between the number of function and gradient evaluations. The only case where unit steps perform substantially better than estimated ones is on f^{inc} with $1 < \lambda_{\mathrm{NOM}} < 3$. These values of λ_{NOM} inflate the Hessian estimates more than BFGS. We suspect that the inflated Hessians are producing estimated steps that are too short. Significantly, estimated steps are *uniformly better* than unit steps on f^{sin}, for which Hessian changes are fairly unpredictable.

6. Discussion. This paper has investigated two estimation problems that arise in the design of quasi-Newton algorithms: (1) estimation of Newton directions by way of sequential updates to a Hessian estimate; and (2) estimation of the optimum along a given search direction. SQN solves the two problems rather differently, using a least-change principle for the Hessian update and a statistical model for the step size. This raises the question of why the statistical model is not also used for the Hessian update.

Straightforward application of the Wishart model leads, in fact, to the BFGS update as is seen in (30). Another derivation of BFGS is obtained by taking the negative logarithm of the Wishart density (26), dividing by $\nu/2$, and taking $\nu \to \infty$. The result is the following function:

$$\psi(\tilde{B}_+) \equiv \operatorname{tr}\left(\tilde{B}_+\right) - \ln|\tilde{B}_+|.$$

Fletcher [16] demonstrated that BFGS minimizes $\psi(\tilde{B}_+) = \psi(E_B + I) = \psi(E_B^* + I)$, where E_B and E_B^* are defined in (13) and (14), respectively. Similarly DFP minimizes $\psi(\tilde{B}_+^{-1}) = \psi((E_B + I)^{-1}) = \psi((E_B^* + I)^{-1})$. Once again, the measure of change is influential.

We argue, however, that accuracy preservation (as measured by E_B^* and E_B) is more important than achieving the best one-step statistical estimate for the problem of *sequentially* estimating the Hessian matrix; this leads to the least-change formulation of Theorem 1. But the SQN update can also be derived from a statistical approach. We first obtained it by combining the Wishart model (25) with a prior distribution that strongly forced C toward the identity matrix. The prior was the statistical embodiment of the least-change principle. Details of this derivation are omitted to save space.

There is a fascinating historical connection that ties the relative change matrices E_B in (13) and E_B^* in (14) to BFGS, DFP, and the E_I method [21] from which Goldfarb [20] derived BFGS. E_W^* and E_W in (11) and (12) are well-known duals that measure change on the inverse and nominal scales and lead to the BFGS and DFP updates, respectively. In the same sense, the dual of E_B is

$$E_I \equiv B_k^{1/2}\left(B_{k+1}^{-1} - B_k^{-1}\right)B_k^{1/2},$$

which is the matrix that Greenstadt minimized. Therefore the SQN update derived from E_B and E_B^* is the dual of the E_I update in the same sense that BFGS is the dual of the older DFP method. Although Greenstadt did not constrain B_{k+1} to be positive definite, minimizing $\|E_I\|_F$ over positive semidefinite updates results in truncating Greenstadt's solution at the critical value of the Broyden parameter. E_B^* was used in [1] to derive an optimally scaled BFGS update. Lukšan [24] generalized the technique to the Broyden family. Specializing Lukšan's result to the case of no scaling produces $\lambda = 0$ for $r < 1$.

Use of a statistical framework to design a quasi-Newton method motivates several interesting topics. The numerical results on three new test functions suggest that information on the bias of previous Hessian estimates could be captured and used to obtain a better update that uses either varying values of λ_{NOM} within the Broyden family or a self-scaling update outside of the Broyden family. Use of the Wishart model to estimate the optimal step size also suggests a more general class of quasi-Newton methods obtained by searching not in the estimated Newton direction $-B^{-1}g$ but rather in an alternate direction determined from the conditional distribution $[-G(x)^{-1}g|a, b]$. We have obtained promising results in some limited tests of these ideas.

Appendix A. Proof of Proposition 1.

Proof. If $r \leq 1$, then $C_{\mathrm{SQN}} = I_{n-1}$ and (23) holds as an equality for *all* u. Suppose $r > 1$ so that $C_{\mathrm{SQN}} = I_{n-1} + (1 - r^{-1})a^{-1}bb'$. Then for any $u : u'b = 0$,

$$u'C_{\mathrm{SQN}}u = u'I_{n-1}u,$$

and thus (23) holds as an equality. Suppose $u = \rho b$ for some $\rho \neq 0$. Positive definiteness of the true Hessian implies $(C_{\mathrm{TRUE}} - a^{-1}bb') > 0$, and thus

$$\begin{aligned} u'C_{\mathrm{TRUE}}u \;>\; a^{-1}u'bb'u &= \rho^2(b'b)r \\ &= u'C_{\mathrm{SQN}}u > \rho^2(b'b) \;=\; u'I_{n-1}u \;>\; 0. \end{aligned}$$

That is, in the direction of u, C_{SQN} is closer to C_{TRUE} than I_{n-1} is, and this implies that (23) holds as a strict inequality.

To prove the final statement, suppose that $C_{\mathrm{TRUE}} = I_{n-1}$ so that the right-hand side of (23) equals zero and consider two cases as follows. First, suppose that $\|V'b\| > 0$ and take $u = \rho b$ with $\rho \neq 0$. Then

$$u'(\hat{C} - C_{\mathrm{TRUE}})u = \rho^2 b' \left(\lambda a^{-1}bb' + VV'\right) b \;> 0,$$

and (23) is violated. On the other hand, if $\|V'b\| = 0$, then assume, without loss of generality, that V has full column rank and take $u = V(V'V)^{-1}y$ for some vector $y \neq 0$. Then $u'b = y'(V'V)^{-1}V'b = 0$ but

$$u'(\hat{C} - C_{\mathrm{TRUE}})u = y'(V'V)^{-1}V'(VV')V(V'V)^{-1}y = y'y \;>\; 0,$$

which violates (23). □

Appendix B. Proof of Proposition 2.

Proof. Using (16), the relation between B_+ and $\tilde{B}_+$ is given by $B_+ = B^{1/2}U\,\tilde{B}_+U'B^{1/2}$. This can be expressed as follows:

$$\begin{aligned} B_+ &= B^{1/2}U \begin{bmatrix} a & b' \\ b & I_{n-1} + \lambda bb'/a \end{bmatrix} U'B^{1/2} \\ &= B + B^{1/2}U \begin{bmatrix} a-1 & b' \\ b & \lambda bb'/a \end{bmatrix} U'B^{1/2} \\ &= B + B^{1/2}U(D_1 + D_2 + D_3)U'B^{1/2}, \end{aligned} \tag{35}$$

where

$$D_1 = \begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix}, \quad D_2 = \begin{bmatrix} a^2/a & b' \\ b & bb'/a \end{bmatrix}, \quad \text{and} \quad D_3 = \begin{bmatrix} 0 & 0 \\ 0 & a(\lambda-1)bb'/a^2 \end{bmatrix}.$$

Denote by $U[,1]$ the first column of U. Then

$$U[,1] = \frac{B^{1/2}\delta}{(\delta'B\delta)^{1/2}}, \quad \begin{bmatrix} a \\ b \end{bmatrix} = \frac{U'B^{-1/2}\gamma}{(\delta'B\delta)^{1/2}},$$

$$a = \frac{\delta'\gamma}{\delta'B\delta}, \quad \text{and} \quad r \equiv \frac{b'b}{a} = \frac{\gamma'B^{-1}\gamma}{\delta'\gamma} - \frac{\delta'\gamma}{\delta'B\delta}.$$

Simple algebraic operations lead to the following equalities:

$$B^{1/2}UD_1U'B^{1/2} = -B^{1/2}U[,1](U[,1])'B^{1/2} = -\frac{B\delta\delta'B}{\delta'B\delta},$$

$$B^{1/2}UD_2U'B^{1/2} = \frac{1}{a}B^{1/2}U \begin{bmatrix} a \\ b \end{bmatrix} [a, b']U'B^{1/2} = \frac{\gamma\gamma'}{\delta'\gamma},$$

and

$$B^{1/2}UD_3U'B^{1/2} = \frac{a(\lambda-1)}{a^2}B^{1/2}U\left(\begin{bmatrix}a\\b\end{bmatrix}-\begin{bmatrix}a\\0\end{bmatrix}\right)\left(\begin{bmatrix}a\\b\end{bmatrix}-\begin{bmatrix}a\\0\end{bmatrix}\right)'U'B^{1/2}$$
$$= (\lambda-1)(\delta'\gamma)\left(\frac{\gamma}{\delta'\gamma}-\frac{B\delta}{\delta'B\delta}\right)\left(\frac{\gamma}{\delta'\gamma}-\frac{B\delta}{\delta'B\delta}\right)'.$$

From these equalities and (35), we see that the expression for B_+ is identical to (4) with $c = (\lambda-1)(\delta'\gamma)$. ∎

Appendix C. Equivalence of change matrices E_B and E_B^*. A scalar measure is required to define the "size" of the matrix E_B defined in (13). Many of the most common scalar measures depend only on eigenvalues—for example, trace, determinant, spectral norm, Frobenius norm, and the ψ-function [9], $\psi(E_B+I) = \operatorname{tr}(E_B+I) - \ln|E_B+I|$.

LEMMA 1. *The eigenvalues of E_B in* (13) *are identical to those of E_B^* in* (14). *Also, the eigenvalues of E_B+I are identical to those of E_B^*+I.*

Proof. Let $\simeq$ denote equality of eigenvalues and note that $PQ \simeq QP$ for square P and Q. Thus,

$$E_B^* \simeq B_{k+1}\left(B_k^{-1}-B_{k+1}^{-1}\right) = \left(B_{k+1}-B_k\right)B_k^{-1} \simeq E_B$$

and

$$E_B^*+I = B_{k+1}^{1/2}B_k^{-1}B_{k+1}^{1/2} \simeq B_k^{-1/2}B_{k+1}B_k^{-1/2} = E_B+I. \quad ∎$$

Acknowledgments. We are grateful to colleagues J. Chambers, D. Gay, D. Lambert, C. Mallows, and M. Wright for helpful discussions. We are also grateful to the reviewers whose careful comments resulted in a clearer presentation.

REFERENCES

[1] M. Al-Baali and R. Fletcher, *Variational methods for nonlinear least-squares*, J. Oper. Res. Soc., 36 (1985), pp. 405–421.
[2] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, 2nd ed., John Wiley and Sons, New York, 1984.
[3] I. Bongartz, A. R. Conn, N. I. M. Gould, and P. L. Toint, *CUTE: Constrained and unconstrained testing environments*, ACM Trans. Math. Software, 21 (1995), pp. 123–160.
[4] C. G. Broyden, *A class of methods for solving nonlinear simultaneous equations*, Math. Comp., 19 (1965), pp. 577–593.
[5] C. G. Broyden, *Quasi-Newton methods and their applications to function minimisation*, Math. Comp., 21 (1967), pp. 368–381.
[6] C. G. Broyden, *The convergence of a class of double rank minimization algorithms: 2. The new algorithm*, J. Inst. Math. Appl., 6 (1970), pp. 222–231.
[7] C. G. Broyden, *On the discovery of the "good Broyden" method*, Math. Program., 87 (2000), pp. 209–213.
[8] R. H. Byrd, D. C. Liu, and J. Nocedal, *On the behavior of Broyden's class of quasi-Newton methods*, SIAM J. Optim., 2 (1992), pp. 533–557.
[9] R. H. Byrd and J. Nocedal, *A tool for the analysis of quasi-Newton methods with application to unconstrained minimization*, SIAM J. Numer. Anal., 26 (1989), pp. 727–739.
[10] R. H. Byrd, J. Nocedal, and Y.-X. Yuan, *Global convergence of a class of quasi-Newton methods on convex problems*, SIAM J. Numer. Anal., 24 (1987), pp. 1171–1190.
[11] W. S. Cleveland, *Robust locally weighted regression and smoothing scatter plots*, J. Amer. Statist. Assoc., 74 (1979), pp. 829–836.

[12] J. B. Crockett and H. Chernoff, *Gradient method of maximization*, Pacific J. Math., 5 (1955), pp. 33–50.
[13] W. C. Davidon, *Variable metric method for minimization*, SIAM J. Optim., 1 (1991), pp. 1–17.
[14] R. Fletcher, *A new approach to variable metric algorithms*, Comput. J., 13 (1970), pp. 317–322.
[15] R. Fletcher, *Practical Methods of Optimization*, 2nd ed., John Wiley and Sons, New York, 1987.
[16] R. Fletcher, *A new variational result for quasi-Newton formulae*, SIAM J. Optim., 1 (1991), pp. 18–21.
[17] R. Fletcher, *An overview of unconstrained optimization*, in Algorithms for Continuous Optimization: The State of the Art, E. Spedicato, ed., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1994, pp. 109–143.
[18] R. Fletcher and M. J. D. Powell, *A rapidly convergent descent method for minimization*, Comput. J., 6 (1963), pp. 163–168.
[19] P. E. Gill and W. Murray, *Performance evaluation for nonlinear optimization*, in Performance Evaluation for Numerical Software, L. D. Fosdick, ed., North–Holland, Amsterdam, 1979, pp. 221–234.
[20] D. Goldfarb, *A family of variable metric methods derived by variational means*, Math. Comp., 24 (1970), pp. 23–26.
[21] J. Greenstadt, *Variations on variable metric methods*, Math. Comp., 24 (1970), pp. 1–22.
[22] J. Greenstadt, *Reminiscences on the development of the variational approach to Davidon's variable-metric method*, Math. Program., 87 (2000), pp. 265–280.
[23] L. Lukšan, *Computational experience with known variable metric updates*, J. Optim. Theory Appl., 83 (1994), pp. 27–47.
[24] L. Lukšan, *Variationally derived scaling and variable metric updates from the preconvex part of the Broyden family*, J. Optim. Theory Appl., 73 (1992), pp. 299–307.
[25] L. Lukšan and E. Spedicato, *Variable metric methods for unconstrained optimization and nonlinear least squares*, J. Comput. Appl. Math., 124 (2000), pp. 61–95.
[26] R. B. Mifflin and J. L. Nazareth, *The least prior deviation quasi-Newton update*, Math. Programming, 65 (1994), pp. 247–261.
[27] J. J. Moré, B. S. Garbow, and K. E. Hillstrom, *Testing unconstrained optimization software*, ACM Trans. Math. Software, 7 (1981), pp. 17–41.
[28] J. J. Moré and D. J. Thuente, *Line search algorithms with guaranteed sufficient decrease*, ACM Trans. Math. Software, 20 (1994), pp. 286–307.
[29] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, New York, 1999.
[30] D. F. Shanno, *Conditioning of quasi-Newton methods for function minimization*, Math. Comp., 24 (1970), pp. 647–650.
[31] Y. Zhang and R. P. Tewarson, *Quasi-Newton algorithms with updates from the preconvex part of Broyden family*, IMA J. Numer. Anal., 8 (1988), pp. 487–509.
[32] Q. Zhao, *Measures for Least Change Secant Methods*, Master's thesis, University of Waterloo, ON, Canada 1992.

SIAM J. OPTIM.
Vol. 18, No. 4, pp. 1286–1309

ON AUGMENTED LAGRANGIAN METHODS WITH GENERAL LOWER-LEVEL CONSTRAINTS*

R. ANDREANI†, E. G. BIRGIN‡, J. M. MARTÍNEZ†, AND M. L. SCHUVERDT†

Abstract. Augmented Lagrangian methods with general lower-level constraints are considered in the present research. These methods are useful when efficient algorithms exist for solving subproblems in which the constraints are only of the lower-level type. Inexact resolution of the lower-level constrained subproblems is considered. Global convergence is proved using the constant positive linear dependence constraint qualification. Conditions for boundedness of the penalty parameters are discussed. The resolution of location problems in which many constraints of the lower-level set are nonlinear is addressed, employing the spectral projected gradient method for solving the subproblems. Problems of this type with more than 3×10^6 variables and 14×10^6 constraints are solved in this way, using moderate computer time. All the codes are available at http://www.ime.usp.br/~egbirgin/tango/.

Key words. nonlinear programming, augmented Lagrangian methods, global convergence, constraint qualifications, numerical experiments

AMS subject classifications. 49M37, 65F05, 65K05, 90C30

DOI. 10.1137/060654797

1. Introduction.

Many practical optimization problems have the form

$$\text{Minimize } f(x) \text{ subject to } x \in \Omega_1 \cap \Omega_2, \tag{1.1}$$

where the constraint set Ω_2 is such that subproblems of type

$$\text{Minimize } F(x) \text{ subject to } x \in \Omega_2 \tag{1.2}$$

are much easier than problems of type (1.1). By this we mean that there exist efficient algorithms for solving (1.2) that cannot be applied to (1.1). In these cases it is natural to address the resolution of (1.1) by means of procedures that allow one to take advantage of methods that solve (1.2). Several examples of this situation may be found in the expanded report [3].

These problems motivated us to revisit augmented Lagrangian methods with arbitrary lower-level constraints. Penalty and augmented Lagrangian algorithms can take advantage of the existence of efficient procedures for solving partially constrained subproblems in a natural way. For this reason, many practitioners in Chemistry, Physics, Economics, and Engineering rely on empirical penalty approaches when they incorporate additional constraints to models that were satisfactorily solved by pre-existing algorithms.

The general structure of augmented Lagrangian methods is well known [7, 22, 39]. An outer iteration consists of two main steps: (a) Minimize the augmented

*Received by the editors March 22, 2006; accepted for publication (in revised form) April 3, 2007; published electronically November 7, 2007. This work was supported by PRONEX-Optimization (PRONEX - CNPq / FAPERJ E-26 / 171.164/2003 - APQ1), FAPESP (grants 2001/04597-4, 2002/00832-1, and 2003/09169-6), and CNPq.

http://www.siam.org/journals/siopt/18-4/65479.html

†Department of Applied Mathematics, IMECC, University of Campinas, CP 6065, 13081-970 Campinas SP, Brazil (andreani@ime.unicamp.br, martinez@ime.unicamp.br, schuverd@ime.unicamp.br).

‡Department of Computer Science IME, University of São Paulo, Rua do Matão 1010, Cidade Universitária, 05508-090, São Paulo SP, Brazil (egbirgin@ime.usp.br).

Lagrangian on the appropriate "simple" set (Ω_2 in our case); (b) update multipliers and penalty parameters. However, several decisions need to be made in order to define a practical algorithm. In this paper we use the Powell–Hestenes–Rockafellar (PHR) augmented Lagrangian function [33, 40, 42] (see [8] for a comparison with other augmented Lagrangian functions), and we keep inequality constraints as they are, instead of replacing them by equality constraints plus bounds. So, we pay the price of having discontinuous second derivatives in the objective function of the subproblems when Ω_1 involves inequalities.

A good criterion is needed for deciding that a suitable approximate subproblem minimizer has been found at step (a). In particular, one must decide whether subproblem minimizers must be feasible with respect to Ω_2 and which is the admissible level of infeasibility and lack of complementarity at these solutions. (Bertsekas [6] analyzed an augmented Lagrangian method for solving (1.1) in the case in which the subproblems are solved exactly.) Moreover, simple and efficient rules for updating multipliers and penalty parameters must be given.

Algorithmic decisions are taken looking at theoretical convergence properties and practical performance. Only experience tells one which theoretical results have practical importance and which do not. Although we recognize that this point is controversial, we would like to make explicit here our own criteria:

1. External penalty methods have the property that, when one finds the *global* minimizers of the subproblems, every limit point is a global minimizer of the original problem [24]. We think that this property must be preserved by the augmented Lagrangian counterparts. This is the main reason why, in our algorithm, we will force boundedness of the Lagrange multipliers estimates.
2. We aim for feasibility of the limit points, but, since this may be impossible (even an empty feasible region is not excluded), a "feasibility result" must say that limit points are stationary points for some infeasibility measure. Some methods require that a constraint qualification hold at all the (feasible or infeasible) iterates. In [15, 47] it was shown that, in such cases, convergence to infeasible points that are not stationary for infeasibility may occur.
3. Feasible limit points that satisfy a constraint qualification must be KKT. The constraint qualification must be as *weak* as possible. Therefore, under the assumption that all the *feasible* points satisfy the constraint qualification, all the feasible limit points should be KKT.
4. Theoretically, it is impossible to prove that the whole sequence generated by a general augmented Lagrangian method converges, because multiple solutions of the subproblems may exist and solutions of the subproblems may oscillate. However, since one uses the solution of one subproblem as the initial point for solving the following one, the convergence of the whole sequence generally occurs. In this case, under suitable local conditions, we must be able to prove that the penalty parameters remain bounded.

In other words, the method must have all the good global convergence properties of an external penalty method. In addition, when everything "goes well," it must be free of the asymptotic instability caused by large penalty parameters. Since we deal with nonconvex problems, the possibility of obtaining full global convergence properties based on proximal-point arguments is out of the question.

The algorithm presented in this paper satisfies those theoretical requirements. In particular, we will show that, if a feasible limit point satisfies the constant positive linear dependence (CPLD) condition, then it is a KKT point. A feasible point x

of a nonlinear programming problem is said to satisfy CPLD if the existence of a nontrivial null linear combination of gradients of active constraints with nonnegative coefficients corresponding to the inequalities implies that the gradients involved in that combination are linearly dependent for all z in a neighborhood of x. The CPLD condition was introduced by Qi and Wei [41]. In [4] it was proved that CPLD is a constraint qualification, being strictly weaker than the linear independence constraint qualification (LICQ) and than the Mangasarian–Fromovitz constraint qualification (MFCQ) [36, 43]. Since CPLD is weaker than (say) LICQ, theoretical results saying that *if a limit point satisfies CPLD then it satisfies KKT* are stronger than theoretical results saying that *if a limit point satisfies LICQ then it satisfies KKT*.

Most practical nonlinear programming methods published after 2001 rely on (a combination of) sequential quadratic programming (SQP), Newton-like, or barrier approaches [1, 5, 14, 16, 18, 19, 26, 27, 28, 29, 35, 38, 44, 45, 46, 48, 49, 50]. None of these methods can be easily adapted to the situation described by (1.1)–(1.2).

In the numerical experiments we will show that, in some very large scale location problems, using a specific algorithm for convex-constrained programming [11, 12, 13, 23] for solving the subproblems in the augmented Lagrangian context is much more efficient than using a general purpose method. We will also show that Algencan (the particular implementation of the algorithm introduced in this paper for the case in which the lower-level set is a box [9]) seems to converge to global minimizers more often than Ipopt [47, 48].

This paper is organized as follows. A high-level description of the main algorithm is given in section 2. The rigorous definition of the method is in section 3. Section 4 is devoted to global convergence results. In section 5 we prove boundedness of the penalty parameters. In section 6 we show the numerical experiments. Conclusions are given in section 7.

Notation. We denote $\mathbb{R}_+ = \{t \in \mathbb{R} \mid t \geq 0\}$, $\mathbb{N} = \{0, 1, 2, \ldots\}$, $\|\cdot\|$ as an arbitrary vector norm, and $[v]_i$ as the ith component of the vector v. If there is no possibility of confusion we may also use the notation v_i. For all $y \in \mathbb{R}^n$, $y_+ = (\max\{0, y_1\}, \ldots, \max\{0, y_n\})^T$. If $F : \mathbb{R}^n \to \mathbb{R}^m$, $F = (f_1, \ldots, f_m)^T$, we denote $\nabla F(x) = (\nabla f_1(x), \ldots, \nabla f_m(x)) \in \mathbb{R}^{n \times m}$. If $K = \{k_0, k_1, k_2, \ldots\} \subset \mathbb{N}$ ($k_{j+1} > k_j$ for all j), we denote $\lim_{k \in K} x_k = \lim_{j \to \infty} x_{k_j}$.

2. Overview of the method. We will consider the following nonlinear programming problem:

$$\text{Minimize } f(x) \text{ subject to } h_1(x) = 0,\ g_1(x) \leq 0,\ h_2(x) = 0,\ g_2(x) \leq 0, \tag{2.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}, h_1 : \mathbb{R}^n \to \mathbb{R}^{m_1}, h_2 : \mathbb{R}^n \to \mathbb{R}^{m_2}, g_1 : \mathbb{R}^n \to \mathbb{R}^{p_1}, g_2 : \mathbb{R}^n \to \mathbb{R}^{p_2}$. We assume that all these functions admit continuous first derivatives on a sufficiently large and open domain. We define $\Omega_1 = \{x \in \mathbb{R}^n \mid h_1(x) = 0, g_1(x) \leq 0\}$ and $\Omega_2 = \{x \in \mathbb{R}^n \mid h_2(x) = 0, g_2(x) \leq 0\}$.

For all $x \in \mathbb{R}^n, \rho > 0, \lambda \in \mathbb{R}^{m_1}, \mu \in \mathbb{R}^{p_1}_+$, we define the augmented Lagrangian with respect to Ω_1 [33, 40, 42] as

$$L(x, \lambda, \mu, \rho) = f(x) + \frac{\rho}{2} \sum_{i=1}^{m_1} \left([h_1(x)]_i + \frac{\lambda_i}{\rho} \right)^2 + \frac{\rho}{2} \sum_{i=1}^{p_1} \left([g_1(x)]_i + \frac{\mu_i}{\rho} \right)_+^2. \tag{2.2}$$

The main algorithm defined in this paper will consist of a sequence of (approximate) minimizations of $L(x, \lambda, \mu, \rho)$ subject to $x \in \Omega_2$, followed by the updating of λ, μ, and ρ. A version of the algorithm with several penalty parameters may be found in [3]. Each approximate minimization of L will be called an *outer iteration*.

After each outer iteration one wishes for some progress in terms of *feasibility* and *complementarity*. The *infeasibility* of x with respect to the equality constraint $[h_1(x)]_i = 0$ is naturally represented by $|[h_1(x)]_i|$. The case of inequality constraints is more complicated because, besides feasibility, one expects to have a null multiplier estimate if $g_i(x) < 0$. A suitable combined measure of infeasibility and non-complementarity with respect to the constraint $[g_1(x)]_i \leq 0$ comes from defining $[\sigma(x, \mu, \rho)]_i = \max\{[g_1(x)]_i, -\mu_i/\rho\}$. Since μ_i/ρ is always nonnegative, it turns out that $[\sigma(x, \mu, \rho)]_i$ vanishes in two situations: (a) when $[g_1(x)]_i = 0$; and (b) when $[g_1(x)]_i < 0$ and $\mu_i = 0$. So, roughly speaking, $|[\sigma(x, \mu, \rho)]_i|$ measures infeasibility and complementarity with respect to the inequality constraint $[g_1(x)]_i \leq 0$. If, between two consecutive outer iterations, enough progress is observed in terms of (at least one of) feasibility and complementarity, the penalty parameter will not be updated. Otherwise, the penalty parameter is increased by a fixed factor.

The rules for updating the multipliers need some discussion. In principle, we adopt the classical first-order correction rule [33, 40, 43], but, in addition, we impose that the multiplier estimates must be bounded. So, we will explicitly project the estimates on a compact box after each update. The reason for this decision was already given in the introduction: we want to preserve the property of external penalty methods that global minimizers of the original problem are obtained if each outer iteration computes a global minimizer of the subproblem. This property is maintained if the quotient of *the square* of each multiplier estimate over the penalty parameter tends to zero when the penalty parameter tends to infinity. We were not able to prove that this condition holds automatically for usual estimates and, in fact, we conjecture that it does not. Therefore, we decided to force the boundedness condition. The price paid for this decision seems to be moderate: in the proof of the boundedness of penalty parameters we will need to assume that the true Lagrange multipliers are within the bounds imposed by the algorithm. Since "large Lagrange multipliers" are a symptom of "near-nonfulfillment" of the MFCQ, this assumption seems to be compatible with the remaining ones that are necessary to prove penalty boundedness.

3. Description of the augmented Lagrangian algorithm. In this section we provide a detailed description of the main algorithm. Approximate solutions of the subproblems are defined as points that satisfy the conditions (3.1)–(3.4) below. These formulae are relaxed KKT conditions of the problem of minimizing L subject to $x \in \Omega_2$. The first-order approximations of the multipliers are computed at Step 3. Lagrange multipliers estimates are denoted λ_k and μ_k, whereas their safeguarded counterparts are $\bar{\lambda}_k$ and $\bar{\mu}_k$. At Step 4 we update the penalty parameters according to the progress in terms of feasibility and complementarity.

Algorithm 3.1. Let $x_0 \in \mathbb{R}^n$ be an arbitrary initial point. The given parameters for the execution of the algorithm are $\tau \in [0, 1), \gamma > 1, \rho_1 > 0$, $-\infty < [\bar{\lambda}_{\min}]_i \leq [\bar{\lambda}_{\max}]_i < \infty$ for all $i = 1, \ldots, m_1$, $0 \leq [\bar{\mu}_{\max}]_i < \infty$ for all $i = 1, \ldots, p_1$, $[\bar{\lambda}_1]_i \in [[\bar{\lambda}_{\min}]_i, [\bar{\lambda}_{\max}]_i]$ for all $i = 1, \ldots, m_1$, $[\bar{\mu}_1]_i \in [0, [\bar{\mu}_{\max}]_i]$ for all $i = 1, \ldots, p_1$. Finally, $\{\varepsilon_k\} \subset \mathbb{R}_+$ is a sequence of tolerance parameters such that $\lim_{k\to\infty} \varepsilon_k = 0$.

Step 1. *Initialization.* Set $k \leftarrow 1$. For $i = 1, \ldots, p_1$, compute $[\sigma_0]_i = \max\{0, [g_1(x_0)]_i\}$.

Step 2. *Solving the subproblem.* Compute (if possible) $x_k \in \mathbb{R}^n$ such that there exist $v_k \in \mathbb{R}^{m_2}, u_k \in \mathbb{R}^{p_2}$ satisfying

$$\left\| \nabla L(x_k, \bar{\lambda}_k, \bar{\mu}_k, \rho_k) + \sum_{i=1}^{m_2} [v_k]_i \nabla [h_2(x_k)]_i + \sum_{i=1}^{p_2} [u_k]_i \nabla [g_2(x_k)]_i \right\| \leq \varepsilon_{k,1}, \tag{3.1}$$

$$[u_k]_i \geq 0 \text{ and } [g_2(x_k)]_i \leq \varepsilon_{k,2} \ \forall \ i = 1, \ldots, p_2, \tag{3.2}$$

$$[g_2(x_k)]_i < -\varepsilon_{k,2} \Rightarrow [u_k]_i = 0 \ \forall \ i = 1, \ldots, p_2, \tag{3.3}$$

$$\|h_2(x_k)\| \leq \varepsilon_{k,3}, \tag{3.4}$$

where $\varepsilon_{k,1}, \varepsilon_{k,2}, \varepsilon_{k,3} \geq 0$ are such that $\max\{\varepsilon_{k,1}, \varepsilon_{k,2}, \varepsilon_{k,3}\} \leq \varepsilon_k$. If it is not possible to find x_k satisfying (3.1)–(3.4), stop the execution of the algorithm.

Step 3. *Estimate multipliers.* For all $i = 1, \ldots, m_1$, compute

$$[\lambda_{k+1}]_i = [\bar{\lambda}_k]_i + \rho_k [h_1(x_k)]_i, \tag{3.5}$$

$$[\bar{\lambda}_{k+1}]_i \in [[\bar{\lambda}_{\min}]_i, [\bar{\lambda}_{\max}]_i]. \tag{3.6}$$

(Usually, $[\bar{\lambda}_{k+1}]_i$ will be the projection of $[\lambda_{k+1}]_i$ on the interval $[[\bar{\lambda}_{\min}]_i, [\bar{\lambda}_{\max}]_i]$.) For all $i = 1, \ldots, p_1$, compute

$$[\mu_{k+1}]_i = \max\{0, [\bar{\mu}_k]_i + \rho_k [g_1(x_k)]_i\}, \quad [\sigma_k]_i = \max\left\{[g_1(x_k)]_i, -\frac{[\bar{\mu}_k]_i}{\rho_k}\right\}, \tag{3.7}$$

$$[\bar{\mu}_{k+1}]_i \in [0, [\bar{\mu}_{\max}]_i].$$

(Usually, $[\bar{\mu}_{k+1}]_i = \min\{[\mu_{k+1}]_i, [\bar{\mu}_{\max}]_i\}$.)

Step 4. *Update the penalty parameter.* If

$$\max\{\|h_1(x_k)\|_\infty, \|\sigma_k\|_\infty\} \leq \tau \max\{\|h_1(x_{k-1})\|_\infty, \|\sigma_{k-1}\|_\infty\},$$

then define $\rho_{k+1} = \rho_k$. Else, define $\rho_{k+1} = \gamma \rho_k$.

Step 5. *Begin a new outer iteration.* Set $k \leftarrow k + 1$. Go to Step 2.

4. Global convergence. In this section we assume that the algorithm does not stop at Step 2. In other words, it is always possible to find x_k satisfying (3.1)–(3.4). Problem-dependent sufficient conditions for this assumption can be given in many cases.

We will also assume that at least a limit point of the sequence generated by Algorithm 3.1 exists. A sufficient condition for this is the existence of $\varepsilon > 0$ such that the set $\{x \in \mathbb{R}^n \mid g_2(x) \leq \varepsilon, \|h_2(x)\| \leq \varepsilon\}$ is bounded. This condition may be enforced, adding artificial simple constraints to the set Ω_2.

Global convergence results that use the CPLD constraint qualification are stronger than previous results for more specific problems: In particular, Conn, Gould, and Toint [21] and Conn et al. [20] proved global convergence of augmented Lagrangian methods for equality constraints and linear constraints, assuming linear independence of all the gradients of active constraints at the limit points. Their assumption is much stronger than our CPLD assumptions. On one hand, the CPLD assumption is weaker than LICQ (for example, CPLD always holds when the constraints are linear). On the other hand, our CPLD assumption involves only feasible points instead of all possible limit points of the algorithm.

Convergence proofs for augmented Lagrangian methods with equalities and box constraints using CPLD were given in [2].

We are going to investigate the status of the limit points of sequences generated by Algorithm 3.1. First, we will prove a result on the feasibility properties of a

limit point. Theorem 4.1 shows that either a limit point is feasible or, if the CPLD constraint qualification with respect to Ω_2 holds, it is a KKT point of the sum of squares of upper-level infeasibilities.

THEOREM 4.1. *Let $\{x_k\}$ be a sequence generated by Algorithm* 3.1. *Let x_* be a limit point of $\{x_k\}$. Then, if the sequence of penalty parameters $\{\rho_k\}$ is bounded, the limit point x_* is feasible. Otherwise, at least one of the following possibilities holds:*

(i) *x_* is a KKT point of the problem*

$$\text{(4.1)}\quad \textit{Minimize } \frac{1}{2}\left[\sum_{i=1}^{m_1}[h_1(x)]_i^2+\sum_{i=1}^{p_1}\max\{0,[g_1(x)]_i\}^2\right] \textit{ subject to } x\in\Omega_2.$$

(ii) *x_* does not satisfy the CPLD constraint qualification associated with Ω_2.*

Proof. Let K be an infinite subsequence in $\mathbb{N}$ such that $\lim_{k\in K} x_k = x_*$. Since $\varepsilon_k \to 0$, by (3.2) and (3.4), we have that $g_2(x_*)\le 0$ and $h_2(x_*)=0$. So, $x_*\in\Omega_2$.

Now, we consider two possibilities: (a) the sequence $\{\rho_k\}$ is bounded; and (b) the sequence $\{\rho_k\}$ is unbounded. Let us analyze first case (a). In this case, from some iteration on, the penalty parameters are not updated. Therefore, $\lim_{k\to\infty}\|h_1(x_k)\| = \lim_{k\to\infty}\|\sigma_k\| = 0$. Thus, $h_1(x_*)=0$. Now, if $[g_1(x_*)]_j > 0$ then $[g_1(x_k)]_j > c > 0$ for $k\in K$ large enough. This would contradict the fact that $[\sigma_k]_j\to 0$. Therefore, $[g_1(x_*)]_i\le 0$ for all $i=1,\dots,p_1$.

Since $x_*\in\Omega_2$, $h_1(x_*)=0$, and $g_1(x_*)\le 0$, x_* is feasible. Therefore, we proved the desired result in the case that $\{\rho_k\}$ is bounded.

Consider now case (b). So, $\{\rho_k\}_{k\in K}$ is not bounded. By (2.2) and (3.1), we have

$$\begin{aligned}\nabla f(x_k)+\sum_{i=1}^{m_1}([\bar\lambda_k]_i+\rho_k[h_1(x_k)]_i)\nabla[h_1(x_k)]_i+\sum_{i=1}^{p_1}\max\{0,[\bar\mu_k]_i\\ +\rho_k[g_1(x_k)]_i\}\nabla[g_1(x_k)]_i+\sum_{i=1}^{m_2}[v_k]_i\nabla[h_2(x_k)]_i+\sum_{j=1}^{p_2}[u_k]_j\nabla[g_2(x_k)]_j=\delta_k,\end{aligned}\tag{4.2}$$

where, since $\varepsilon_k\to 0$, $\lim_{k\in K}\|\delta_k\|=0$.

If $[g_2(x_*)]_i<0$, there exists $k_1\in\mathbb{N}$ such that $[g_2(x_k)]_i<-\varepsilon_k$ for all $k\ge k_1, k\in K$. Therefore, by (3.3), $[u_k]_i=0$ for all $k\in K, k\ge k_1$. Thus, by $x_*\in\Omega_2$ and (4.2), for all $k\in K, k\ge k_1$ we have that

$$\begin{aligned}\nabla f(x_k)+\sum_{i=1}^{m_1}([\bar\lambda_k]_i+\rho_k[h_1(x_k)]_i)\nabla[h_1(x_k)]_i+\sum_{i=1}^{p_1}\max\{0,[\bar\mu_k]_i\\ +\rho_k[g_1(x_k)]_i\}\nabla[g_1(x_k)]_i+\sum_{i=1}^{m_2}[v_k]_i\nabla[h_2(x_k)]_i+\sum_{[g_2(x_*)]_j=0}[u_k]_j\nabla[g_2(x_k)]_j=\delta_k.\end{aligned}$$

Dividing by ρ_k we get

$$\begin{aligned}\frac{\nabla f(x_k)}{\rho_k} &+ \sum_{i=1}^{m_1}\left(\frac{[\bar{\lambda}_k]_i}{\rho_k} + [h_1(x_k)]_i\right)\nabla[h_1(x_k)]_i \\ &+ \sum_{i=1}^{p_1}\max\left\{0, \frac{[\bar{\mu}_k]_i}{\rho_k} + [g_1(x_k)]_i\right\}\nabla[g_1(x_k)]_i \\ &+ \sum_{i=1}^{m_2}\frac{[v_k]_i}{\rho_k}\nabla[h_2(x_k)]_i \\ &+ \sum_{[g_2(x_*)]_j=0}\frac{[u_k]_j}{\rho_k}\nabla[g_2(x_k)]_j = \frac{\delta_k}{\rho_k}.\end{aligned}$$

By Caratheodory's theorem of cones (see [7, page 689]) there exist $\widehat{I}_k \subset \{1,\ldots,m_2\}$, $\widehat{J}_k \subset \{j \mid [g_2(x_*)]_j = 0\}$, $[\widehat{v}_k]_i$, $i \in \widehat{I}_k$, and $[\widehat{u}_k]_j \geq 0$, $j \in \widehat{J}_k$, such that the vectors $\{\nabla[h_2(x_k)]_i\}_{i\in\widehat{I}_k} \cup \{\nabla[g_2(x_k)]_j\}_{j\in\widehat{J}_k}$ are linearly independent and

$$\begin{aligned}\frac{\nabla f(x_k)}{\rho_k} &+ \sum_{i=1}^{m_1}\left(\frac{[\bar{\lambda}_k]_i}{\rho_k} + [h_1(x_k)]_i\right)\nabla[h_1(x_k)]_i \\ &+ \sum_{i=1}^{p_1}\max\left\{0, \frac{[\bar{\mu}_k]_i}{\rho_k} + [g_1(x_k)]_i\right\}\nabla[g_1(x_k)]_i \\ &+ \sum_{i\in\widehat{I}_k}[\widehat{v}_k]_i\nabla[h_2(x_k)]_i \\ &+ \sum_{j\in\widehat{J}_k}[\widehat{u}_k]_j\nabla[g_2(x_k)]_j = \frac{\delta_k}{\rho_k}.\end{aligned} \tag{4.3}$$

Since there exists a finite number of possible sets $\widehat{I}_k, \widehat{J}_k$, there exists an infinite set of indices K_1 such that $K_1 \subset \{k \in K \mid k \geq k_1\}$, $\widehat{I}_k = \widehat{I}$, and

$$\widehat{J} = \widehat{J}_k \subset \{j \mid [g_2(x_*)]_j = 0\} \tag{4.4}$$

for all $k \in K_1$. Then, by (4.3), for all $k \in K_1$ we have

$$\begin{aligned}\frac{\nabla f(x_k)}{\rho_k} &+ \sum_{i=1}^{m_1}\left(\frac{[\bar{\lambda}_k]_i}{\rho_k} + [h_1(x_k)]_i\right)\nabla[h_1(x_k)]_i \\ &+ \sum_{i=1}^{p_1}\max\left\{0, \frac{[\bar{\mu}_k]_i}{\rho_k} + [g_1(x_k)]_i\right\}\nabla[g_1(x_k)]_i \\ &+ \sum_{i\in\widehat{I}}[\widehat{v}_k]_i\nabla[h_2(x_k)]_i \\ &+ \sum_{j\in\widehat{J}}[\widehat{u}_k]_j\nabla[g_2(x_k)]_j = \frac{\delta_k}{\rho_k},\end{aligned} \tag{4.5}$$

and the gradients

$$\{\nabla[h_2(x_k)]_i\}_{i\in\widehat{I}} \cup \{\nabla[g_2(x_k)]_j\}_{j\in\widehat{J}} \tag{4.6}$$

are linearly independent.

We consider again two cases: (1) the sequence $\{\|(\widehat{v}_k, \widehat{u}_k)\|, k \in K_1\}$ is bounded; and (2) the sequence $\{\|(\widehat{v}_k, \widehat{u}_k)\|, k \in K_1\}$ is unbounded. If the sequence $\{\|(\widehat{v}_k, \widehat{u}_k)\|\}_{k\in K_1}$ is bounded, and $\widehat{I} \cup \widehat{J} \neq \emptyset$, there exist $(\widehat{v}, \widehat{u}), \widehat{u} \geq 0$, and an infinite set of indices $K_2 \subset K_1$ such that $\lim_{k\in K_2}(\widehat{v}_k, \widehat{u}_k) = (\widehat{v}, \widehat{u})$. Since $\{\rho_k\}$ is unbounded, by the boundedness of $\bar{\lambda}_k$ and $\bar{\mu}_k$, $\lim[\bar{\lambda}_k]_i/\rho_k = 0 = \lim[\bar{\mu}_k]_j/\rho_k$ for all i, j. Therefore, by $\delta_k \to 0$, taking limits for $k \in K_2$ in (4.5), we obtain

$$\begin{aligned}\sum_{i=1}^{m_1}[h_1(x_*)]_i\nabla[h_1(x_*)]_i + \sum_{i=1}^{p_1}\max\{0, [g_1(x_*)]_i\}\nabla[g_1(x_*)]_i \\ + \sum_{i\in\widehat{I}}\widehat{v}_i\nabla[h_2(x_*)]_i + \sum_{j\in\widehat{J}}\widehat{u}_j\nabla[g_2(x_*)]_j = 0.\end{aligned} \tag{4.7}$$

If $\widehat{I} \cup \widehat{J} = \emptyset$ we obtain $\sum_{i=1}^{m_1}[h_1(x_*)]_i\nabla[h_1(x_*)]_i + \sum_{i=1}^{p_1}\max\{0, [g_1(x_*)]_i\}\nabla[g_1(x_*)]_i = 0$.

Therefore, by $x_* \in \Omega_2$ and (4.4), x_* is a KKT point of (4.1).

Finally, assume that $\{\|(\widehat{v}_k, \widehat{u}_k)\|\}_{k\in K_1}$ is unbounded. Let $K_3 \subset K_1$ be such that $\lim_{k\in K_3}\|(\widehat{v}_k, \widehat{u}_k)\| = \infty$ and $(\widehat{v}, \widehat{u}) \neq 0, \widehat{u} \geq 0$, such that $\lim_{k\in K_3}\frac{(\widehat{v}_k, \widehat{u}_k)}{\|(\widehat{v}_k, \widehat{u}_k)\|} = (\widehat{v}, \widehat{u})$. Dividing both sides of (4.5) by $\|(\widehat{v}_k, \widehat{u}_k)\|$ and taking limits for $k \in K_3$, we deduce that $\sum_{i\in\widehat{I}}\widehat{v}_i\nabla[h_2(x_*)]_i + \sum_{j\in\widehat{J}}\widehat{u}_j\nabla[g_2(x_*)]_j = 0$. But $[g_2(x_*)]_j = 0$ for all $j \in \widehat{J}$. Then, by (4.6), x_* does not satisfy the CPLD constraint qualification associated with the set Ω_2. This completes the proof. ∎

Roughly speaking, Theorem 4.1 says that, if x_* is not feasible, then (very likely) it is a local minimizer of the upper-level infeasibility, subject to lower-level feasibility. From the point of view of optimality, we are interested in the status of feasible limit points. In Theorem 4.2 we will prove that, under the CPLD constraint qualification, feasible limit points are stationary (KKT) points of the original problem. Since CPLD is strictly weaker than the MFCQ, it turns out that the following theorem is stronger than results where KKT conditions are proved under MFCQ or regularity assumptions.

THEOREM 4.2. *Let $\{x_k\}_{k\in\mathbb{N}}$ be a sequence generated by Algorithm* 3.1. *Assume that $x_* \in \Omega_1 \cap \Omega_2$ is a limit point that satisfies the CPLD constraint qualification related to $\Omega_1 \cap \Omega_2$. Then, x_* is a KKT point of the original problem* (2.1). *Moreover, if x_* satisfies the MFCQ and $\{x_k\}_{k\in K}$ is a subsequence that converges to x_*, the set*

$$\{\|\lambda_{k+1}\|, \|\mu_{k+1}\|, \|v_k\|, \|u_k\|\}_{k\in K} \text{ is bounded.} \tag{4.8}$$

Proof. For all $k \in \mathbb{N}$, by (3.1), (3.3), (3.5), and (3.7), there exist $u_k \in \mathbb{R}^{p_2}_+$, $\delta_k \in \mathbb{R}^n$ such that $\|\delta_k\| \leq \varepsilon_k$ and

$$\begin{aligned}\nabla f(x_k) + \sum_{i=1}^{m_1}[\lambda_{k+1}]_i\nabla[h_1(x_k)]_i + \sum_{i=1}^{p_1}[\mu_{k+1}]_i\nabla[g_1(x_k)]_i \\ + \sum_{i=1}^{m_2}[v_k]_i\nabla[h_2(x_k)]_i + \sum_{j=1}^{p_2}[u_k]_j\nabla[g_2(x_k)]_j = \delta_k.\end{aligned} \tag{4.9}$$

By (3.7), $\mu_{k+1} \in \mathbb{R}^{p_1}_+$ for all $k \in \mathbb{N}$. Let $K \subset \mathbb{N}$ be such that $\lim_{k\in K} x_k = x_*$. Suppose that $[g_2(x_*)]_i < 0$. Then, there exists $k_1 \in \mathbb{N}$ such that for all $k \in K, k \geq k_1, [g_2(x_k)]_i < -\varepsilon_k$. Then, by (3.3), $[u_k]_i = 0$ for all $k \in K, k \geq k_1$.

Let us prove now that a similar property takes place when $[g_1(x_*)]_i < 0$. In this case, there exists $k_2 \geq k_1$ such that $[g_1(x_k)]_i < c < 0$ for all $k \in K, k \geq k_2$.

We consider two cases: (1) $\{\rho_k\}$ is unbounded; and (2) $\{\rho_k\}$ is bounded. In the first case we have that $\lim_{k\in K} \rho_k = \infty$. Since $\{[\bar{\mu}_k]_i\}$ is bounded, there exists $k_3 \geq k_2$ such that, for all $k \in K, k \geq k_3$, $[\bar{\mu}_k]_i + \rho_k[g_1(x_k)]_i < 0$. By the definition of μ_{k+1} this implies that $[\mu_{k+1}]_i = 0$ for all $k \in K, k \geq k_3$.

Consider now the case in which $\{\rho_k\}$ is bounded. In this case, $\lim_{k\to\infty}[\sigma_k]_i = 0$. Therefore, since $[g_1(x_k)]_i < c < 0$ for $k \in K$ large enough, $\lim_{k\in K}[\bar{\mu}_k]_i = 0$. So, for $k \in K$ large enough, $[\bar{\mu}_k]_i + \rho_k[g_1(x_k)]_i < 0$. By the definition of μ_{k+1}, there exists $k_4 \geq k_2$ such that $[\mu_{k+1}]_i = 0$ for $k \in K, k \geq k_4$.

Therefore, there exists $k_5 \geq \max\{k_1, k_3, k_4\}$ such that for all $k \in K$, $k \geq k_5$,

$$[[g_1(x_*)]_i < 0 \Rightarrow [\mu_{k+1}]_i = 0] \text{ and } [[g_2(x_*)]_i < 0 \Rightarrow [u_k]_i = 0]. \tag{4.10}$$

(Observe that, up to now, we did not use the CPLD condition.) By (4.9) and (4.10), for all $k \in K, k \geq k_5$, we have

$$\begin{aligned} \nabla f(x_k) + \sum_{i=1}^{m_1} [\lambda_{k+1}]_i \nabla [h_1(x_k)]_i + \sum_{[g_1(x_*)]_i=0} [\mu_{k+1}]_i \nabla [g_1(x_k)]_i \\ + \sum_{i=1}^{m_2} [v_k]_i \nabla [h_2(x_k)]_i + \sum_{[g_2(x_*)]_j=0} [u_k]_j \nabla [g_2(x_k)]_j = \delta_k, \end{aligned} \tag{4.11}$$

with $\mu_{k+1} \in \mathbb{R}^{p_1}_+, u_k \in \mathbb{R}^{p_2}_+$.

By Caratheodory's theorem of cones, for all $k \in K, k \geq k_5$, there exist

$$\widehat{I}_k \subset \{1, \ldots, m_1\}, \widehat{J}_k \subset \{j \mid [g_1(x_*)]_j = 0\}, \hat{I}_k \subset \{1, \ldots, m_2\}, \hat{J}_k \subset \{j \mid [g_2(x_*)]_j = 0\},$$

$$[\widehat{\lambda}_k]_i \in \mathbb{R}\ \forall i \in \widehat{I}_k, [\widehat{\mu}_k]_j \geq 0\ \forall j \in \widehat{J}_k, [\widehat{v}_k]_i \in \mathbb{R}\ \forall i \in \hat{I}_k, [\widehat{u}_k]_j \geq 0\ \forall j \in \hat{J}_k$$

such that the vectors

$$\{\nabla [h_1(x_k)]_i\}_{i\in\widehat{I}_k} \cup \{\nabla [g_1(x_k)]_i\}_{i\in\widehat{J}_k} \cup \{\nabla [h_2(x_k)]_i\}_{i\in\hat{I}_k} \cup \{\nabla [g_2(x_k)]_i\}_{i\in\hat{J}_k}$$

are linearly independent and

$$\begin{aligned} \nabla f(x_k) + \sum_{i\in\widehat{I}_k} [\widehat{\lambda}_k]_i \nabla [h_1(x_k)]_i + \sum_{i\in\widehat{J}_k} [\widehat{\mu}_k]_i \nabla [g_1(x_k)]_i \\ + \sum_{i\in\hat{I}_k} [\widehat{v}_k]_i \nabla [h_2(x_k)]_i + \sum_{j\in\hat{J}_k} [\widehat{u}_k]_j \nabla [g_2(x_k)]_j = \delta_k. \end{aligned} \tag{4.12}$$

Since the number of possible sets of indices $\widehat{I}_k, \widehat{J}_k, \hat{I}_k, \hat{J}_k$ is finite, there exists an infinite set $K_1 \subset \{k \in K \mid k \geq k_5\}$ such that $\widehat{I}_k = \widehat{I}, \widehat{J}_k = \widehat{J}, \hat{I}_k = \hat{I}, \hat{J}_k = \hat{J}$ for all $k \in K_1$.

Then, by (4.12),

$$\begin{aligned} \nabla f(x_k) + \sum_{i\in\widehat{I}} [\widehat{\lambda}_k]_i \nabla [h_1(x_k)]_i + \sum_{i\in\widehat{J}} [\widehat{\mu}_k]_i \nabla [g_1(x_k)]_i \\ + \sum_{i\in\hat{I}} [\widehat{v}_k]_i \nabla [h_2(x_k)]_i + \sum_{j\in\hat{J}} [\widehat{u}_k]_j \nabla [g_2(x_k)]_j = \delta_k \end{aligned} \tag{4.13}$$

and the vectors

$$\{\nabla[h_1(x_k)]_i\}_{i\in\widehat{I}} \cup \{\nabla[g_1(x_k)]_i\}_{i\in\widehat{J}} \cup \{\nabla[h_2(x_k)]_i\}_{i\in\hat{I}} \cup \{\nabla[g_2(x_k)]_i\}_{i\in\hat{J}} \tag{4.14}$$

are linearly independent for all $k \in K_1$.

If $\widehat{I} \cup \widehat{J} \cup \hat{I} \cup \hat{J} = \emptyset$, by (4.13) and $\delta_k \to 0$ we obtain $\nabla f(x_*) = 0$. Otherwise, let us define

$$S_k = \max\{\max\{|[\widehat{\lambda}_k]_i|, i \in \widehat{I}\}, \max\{[\widehat{\mu}_k]_i, i \in \widehat{J}\}, \max\{|[\widehat{v}_k]_i|, i \in \hat{I}\}, \max\{[\widehat{u}_k]_i, i \in \hat{J}\}\}.$$

We consider two possibilities: (a) $\{S_k\}_{k\in K_1}$ has a bounded subsequence; and (b) $\lim_{k\in K_1} S_k = \infty$. If $\{S_k\}_{k\in K_1}$ has a bounded subsequence, there exists $K_2 \subset K_1$ such that $\lim_{k\in K_2}[\widehat{\lambda}_k]_i = \widehat{\lambda}_i$, $\lim_{k\in K_2}[\widehat{\mu}_k]_i = \widehat{\mu}_i \geq 0$, $\lim_{k\in K_2}[\widehat{v}_k]_i = \widehat{v}_i$, and $\lim_{k\in K_2}[\widehat{u}_k]_i = \widehat{u}_i \geq 0$. By $\varepsilon_k \to 0$ and $\|\delta_k\| \leq \varepsilon_k$, taking limits in (4.13) for $k \in K_2$, we obtain

$$\nabla f(x_*) + \sum_{i\in\widehat{I}} \widehat{\lambda}_i \nabla[h_1(x_*)]_i + \sum_{i\in\widehat{J}} \widehat{\mu}_i \nabla[g_1(x_*)]_i + \sum_{i\in\hat{I}} \widehat{v}_i \nabla[h_2(x_*)]_i + \sum_{j\in\hat{J}} \widehat{u}_j \nabla[g_2(x_*)]_j = 0,$$

with $\widehat{\mu}_i \geq 0, \widehat{u}_i \geq 0$. Since $x_* \in \Omega_1 \cap \Omega_2$, we have that x_* is a KKT point of (2.1).

Suppose now that $\lim_{k\in K_2} S_k = \infty$. Dividing both sides of (4.13) by S_k we obtain

$$\begin{aligned} &\frac{\nabla f(x_k)}{S_k} + \sum_{i\in\widehat{I}} \frac{[\widehat{\lambda}_k]_i}{S_k} \nabla[h_1(x_k)]_i + \sum_{i\in\widehat{J}} \frac{[\widehat{\mu}_k]_i}{S_k} \nabla[g_1(x_k)]_i \\ &+ \sum_{i\in\hat{I}} \frac{[\widehat{v}_k]_i}{S_k} \nabla[h_2(x_k)]_i + \sum_{j\in\hat{J}} \frac{[\widehat{u}_k]_j}{S_k} \nabla[g_2(x_k)]_j = \frac{\delta_k}{S_k}, \end{aligned} \tag{4.15}$$

where $|\frac{[\widehat{\lambda}_k]_i}{S_k}| \leq 1, |\frac{[\widehat{\mu}_k]_i}{S_k}| \leq 1, |\frac{[\widehat{v}_k]_i}{S_k}| \leq 1, |\frac{[\widehat{u}_k]_j}{S_k}| \leq 1$. Therefore, there exists $K_3 \subset K_1$ such that $\lim_{k\in K_3} \frac{[\widehat{\lambda}_k]_i}{S_k} = \widehat{\lambda}_i, \lim_{k\in K_3} \frac{[\widehat{\mu}_k]_i}{S_k} = \widehat{\mu}_i \geq 0, \lim_{k\in K_3} \frac{[\widehat{v}_k]_i}{S_k} = \widehat{v}_i, \lim_{k\in K_3} \frac{[\widehat{u}_k]_j}{S_k} = \widehat{u}_j \geq 0$. Taking limits on both sides of (4.15) for $k \in K_3$, we obtain

$$\sum_{i\in\widehat{I}} \widehat{\lambda}_i \nabla[h_1(x_*)]_i + \sum_{i\in\widehat{J}} \widehat{\mu}_i \nabla[g_1(x_*)]_i + \sum_{i\in\hat{I}} \widehat{v}_i \nabla[h_2(x_*)]_i + \sum_{j\in\hat{J}} \widehat{u}_j \nabla[g_2(x_*)]_j = 0.$$

But the modulus of at least one of the coefficients $\widehat{\lambda}_i, \widehat{\mu}_i, \widehat{v}_i, \widehat{u}_i$ is equal to 1. Then, by the CPLD condition, the gradients

$$\{\nabla[h_1(x)]_i\}_{i\in\widehat{I}} \cup \{\nabla[g_1(x)]_i\}_{i\in\widehat{J}} \cup \{\nabla[h_2(x)]_i\}_{i\in\hat{I}} \cup \{\nabla[g_2(x)]_i\}_{i\in\hat{J}}$$

must be linearly dependent in a neighborhood of x_*. This contradicts (4.14). Therefore, the main part of the theorem is proved.

Finally, let us prove that the property (4.8) holds if x_* satisfies the MFCQ. Let us define

$$B_k = \max\{\|\lambda_{k+1}\|_\infty, \|\mu_{k+1}\|_\infty, \|v_k\|_\infty, \|u_k\|_\infty\}_{k\in K}.$$

If (4.8) is not true, we have that $\lim_{k\in K} B_k = \infty$. In this case, dividing both sides of (4.11) by B_k and taking limits for an appropriate subsequence, we obtain that x_* does not satisfy the MFCQ. ∎

5. Boundedness of the penalty parameters. When the penalty parameters associated with penalty or augmented Lagrangian methods are too large, the subproblems tend to be ill-conditioned and their resolution becomes harder. One of the main motivations for the development of the basic augmented Lagrangian algorithm is the necessity of overcoming this difficulty. Therefore, the study of conditions under which penalty parameters are bounded plays an important role in augmented Lagrangian approaches.

5.1. Equality constraints. We will consider first the case $p_1 = p_2 = 0$.

Let $f : \mathbb{R}^n \to \mathbb{R}, h_1 : \mathbb{R}^n \to \mathbb{R}^{m_1}, h_2 : \mathbb{R}^n \to \mathbb{R}^{m_2}$. We address the problem

$$\text{Minimize } f(x) \text{ subject to } h_1(x) = 0,\ h_2(x) = 0. \tag{5.1}$$

The Lagrangian function associated with problem (5.1) is given by $L_0(x, \lambda, v) = f(x) + \langle h_1(x), \lambda\rangle + \langle h_2(x), v\rangle$ for all $x \in \mathbb{R}^n$, $\lambda \in \mathbb{R}^{m_1}$, $v \in \mathbb{R}^{m_2}$.

Algorithm 3.1 will be considered with the following standard definition for the safeguarded Lagrange multipliers.

DEFINITION. *For all $k \in \mathbb{N}$, $i = 1, \ldots, m_1$, $[\bar\lambda_{k+1}]_i$ will be the projection of $[\lambda_{k+1}]_i$ on the interval $[[\bar\lambda_{\min}]_i, [\bar\lambda_{\max}]_i]$.*

We will use the following assumptions.

Assumption 1. The sequence $\{x_k\}$ is generated by the application of Algorithm 3.1 to problem (5.1) and $\lim_{k\to\infty} x_k = x_*$.

Assumption 2. The point x_* is feasible ($h_1(x_*) = 0$ and $h_2(x_*) = 0$).

Assumption 3. The gradients $\nabla[h_1(x_*)]_1, \ldots, \nabla[h_1(x_*)]_{m_1}, \nabla[h_2(x_*)]_1, \ldots,$ and $\nabla[h_2(x_*)]_{m_2}$ are linearly independent.

Assumption 4. The functions f, h_1, and h_2 admit continuous second derivatives in a neighborhood of x_*.

Assumption 5. The second-order sufficient condition for local minimizers [25, page 211], holds with Lagrange multipliers $\lambda_* \in \mathbb{R}^{m_1}$ and $v_* \in \mathbb{R}^{m_2}$.

Assumption 6. For all $i = 1, \ldots, m_1$, $[\lambda_*]_i \in ([\bar\lambda_{\min}]_i, [\bar\lambda_{\max}]_i)$.

PROPOSITION 5.1. *Suppose that Assumptions* 1, 2, 3, *and* 6 *hold. Then,* $\lim_{k\to\infty} \lambda_k = \lambda_*$, $\lim_{k\to\infty} v_k = v_*$, *and $\bar\lambda_k = \lambda_k$ for k large enough.*

Proof. The proof of the first part follows from the definition of λ_{k+1}, the stopping criterion of the subproblems, and the linear independence of the gradients of the constraints at x_*. The second part of the thesis is a consequence of $\lambda_k \to \lambda_*$, using Assumption 6 and the definition of $\bar\lambda_{k+1}$. ∎

LEMMA 5.2. *Suppose that Assumptions* 3 *and* 5 *hold. Then, there exists $\bar\rho > 0$ such that, for all $\pi \in [0, 1/\bar\rho]$, the matrix*

$$\begin{pmatrix} \nabla^2_{xx} L_0(x_*, \lambda_*, v_*) & \nabla h_1(x_*) & \nabla h_2(x_*) \\ \nabla h_1(x_*)^T & -\pi I & 0 \\ \nabla h_2(x_*)^T & 0 & 0 \end{pmatrix}$$

is nonsingular.

Proof. The matrix is trivially nonsingular for $\pi = 0$. So, the thesis follows by the continuity of the matricial inverse. ∎

LEMMA 5.3. *Suppose that Assumptions* 1–5 *hold. Let $\bar\rho$ be as in Lemma* 5.2. *Suppose that there exists $k_0 \in \mathbb{N}$ such that $\rho_k \geq \bar\rho$ for all $k \geq k_0$. Define*

$$\alpha_k = \nabla L(x_k, \bar\lambda_k, \rho_k) + \nabla h_2(x_k) v_k, \tag{5.2}$$

$$\beta_k = h_2(x_k). \tag{5.3}$$

Then, there exists $M > 0$ such that, for all $k \in \mathbb{N}$,

$$\|x_k - x_*\| \le M \max\left\{\frac{\|\bar\lambda_k - \lambda_*\|_\infty}{\rho_k}, \|\alpha_k\|, \|\beta_k\|\right\}, \tag{5.4}$$

$$\|\lambda_{k+1} - \lambda_*\| \le M \max\left\{\frac{\|\bar\lambda_k - \lambda_*\|_\infty}{\rho_k}, \|\alpha_k\|, \|\beta_k\|\right\}. \tag{5.5}$$

Proof. Define, for all $k \in \mathbb{N}$,

$$t_k = (\bar\lambda_k - \lambda_*)/\rho_k, \tag{5.6}$$

$$\pi_k = 1/\rho_k. \tag{5.7}$$

By (3.5), (5.2), and (5.3), $\nabla L(x_k, \bar\lambda_k, \rho_k) + \nabla h_2(x_k)v_k - \alpha_k = 0$, $\lambda_{k+1} = \bar\lambda_k + \rho_k h_1(x_k)$, and $h_2(x_k) - \beta_k = 0$ for all $k \in \mathbb{N}$.

Therefore, by (5.6) and (5.7), we have that $\nabla f(x_k) + \nabla h_1(x_k)\lambda_{k+1} + \nabla h_2(x_k)v_k - \alpha_k = 0$, $h_1(x_k) - \pi_k\lambda_{k+1} + t_k + \pi_k\lambda_* = 0$, and $h_2(x_k) - \beta_k = 0$ for all $k \in \mathbb{N}$. Define, for all $\pi \in [0, 1/\bar\rho]$, $F_\pi : \mathbb{R}^n \times \mathbb{R}^{m_1} \times \mathbb{R}^{m_2} \times \mathbb{R}^{m_1} \times \mathbb{R}^n \times \mathbb{R}^{m_2} \to \mathbb{R}^n \times \mathbb{R}^{m_1} \times \mathbb{R}^{m_2}$ by

$$F_\pi(x, \lambda, v, t, \alpha, \beta) = \begin{pmatrix} \nabla f(x) + \nabla h_1(x)\lambda + \nabla h_2(x)v - \alpha \\ [h_1(x)]_1 - \pi[\lambda]_1 + [t]_1 + \pi[\lambda_*]_1 \\ \cdot \\ \cdot \\ \cdot \\ [h_1(x)]_{m_1} - \pi[\lambda]_{m_1} + [t]_{m_1} + \pi[\lambda_*]_{m_1} \\ h_2(x) - \beta \end{pmatrix}.$$

Clearly,

$$F_{\pi_k}(x_k, \lambda_{k+1}, v_k, t_k, \alpha_k, \beta_k) = 0 \tag{5.8}$$

and, by Assumptions 1 and 2,

$$F_\pi(x_*, \lambda_*, v_*, 0, 0, 0) = 0 \ \ \forall \pi \in [0, 1/\bar\rho]. \tag{5.9}$$

Moreover, the Jacobian matrix of F_π with respect to (x, λ, v) computed at $(x_*, \lambda_*, v_*, 0, 0, 0)$ is

$$\begin{pmatrix} \nabla^2_{xx} L_0(x_*, \lambda_*, v_*) & \nabla h_1(x_*) & \nabla h_2(x_*) \\ \nabla h_1(x_*)^T & -\pi I & 0 \\ \nabla h_2(x_*)^T & 0 & 0 \end{pmatrix}.$$

By Lemma 5.2, this matrix is nonsingular for all $\pi \in [0, 1/\bar\rho]$. By continuity, the norm of its inverse is bounded in a neighborhood of $(x_*, \lambda_*, v_*, 0, 0, 0)$ uniformly with respect to $\pi \in [0, 1/\bar\rho]$. Moreover, the first and second derivatives of F_π are also bounded in a neighborhood of $(x_*, \lambda_*, v_*, 0, 0, 0)$ uniformly with respect to $\pi \in [0, 1/\bar\rho]$. Therefore, the bounds (5.4) and (5.5) follow from (5.8) and (5.9) by the implicit function theorem and the mean value theorem of integral calculus. ∎

THEOREM 5.4. *Suppose that Assumptions* 1–6 *are satisfied by the sequence generated by Algorithm* 3.1 *applied to the problem* (5.1). *In addition, assume that there*

exists a sequence $\eta_k \to 0$ *such that* $\varepsilon_k \leq \eta_k \|h_1(x_k)\|_\infty$ *for all* $k \in \mathbb{N}$. *Then, the sequence of penalty parameters* $\{\rho_k\}$ *is bounded.*

Proof. Assume, by contradiction, that $\lim_{k\to\infty} \rho_k = \infty$. Since $h_1(x_*) = 0$, by the continuity of the first derivatives of h_1 there exists $L > 0$ such that, for all $k \in \mathbb{N}$, $\|h_1(x_k)\|_\infty \leq L\|x_k - x_*\|$. Therefore, by the hypothesis, (5.4), and Proposition 5.1, we have that $\|h_1(x_k)\|_\infty \leq LM \max\{\frac{\|\lambda_k - \lambda_*\|_\infty}{\rho_k}, \eta_k \|h_1(x_k)\|_\infty\}$ for k large enough. Since η_k tends to zero, this implies that

$$\|h_1(x_k)\|_\infty \leq LM \frac{\|\lambda_k - \lambda_*\|_\infty}{\rho_k} \tag{5.10}$$

for k large enough.

By (3.6) and Proposition 5.1, we have that $\lambda_k = \lambda_{k-1} + \rho_{k-1} h_1(x_{k-1})$ for k large enough. Therefore,

$$\|h_1(x_{k-1})\|_\infty = \frac{\|\lambda_k - \lambda_{k-1}\|_\infty}{\rho_{k-1}} \geq \frac{\|\lambda_{k-1} - \lambda_*\|_\infty}{\rho_{k-1}} - \frac{\|\lambda_k - \lambda_*\|_\infty}{\rho_{k-1}}. \tag{5.11}$$

Now, by (5.5), the hypothesis of this theorem, and Proposition 5.1, for k large enough we have $\|\lambda_k - \lambda_*\|_\infty \leq M(\frac{\|\lambda_{k-1} - \lambda_*\|_\infty}{\rho_{k-1}} + \eta_{k-1}\|h_1(x_{k-1})\|_\infty)$. This implies that $\frac{\|\lambda_{k-1} - \lambda_*\|_\infty}{\rho_{k-1}} \geq \frac{\|\lambda_k - \lambda_*\|_\infty}{M} - \eta_{k-1}\|h_1(x_{k-1})\|_\infty$. Therefore, by (5.11), $(1 + \eta_{k-1})$ $\|h_1(x_{k-1})\|_\infty \geq \|\lambda_k - \lambda_*\|_\infty(\frac{1}{M} - \frac{1}{\rho_{k-1}}) \geq \frac{1}{2M}\|\lambda_k - \lambda_*\|_\infty$. Thus, $\|\lambda_k - \lambda_*\|_\infty \leq 3M\|h_1(x_{k-1})\|_\infty$ for k large enough. By (5.10), we have that $\|h_1(x_k)\|_\infty \leq \frac{3LM^2}{\rho_k}\|h_1(x_{k-1})\|_\infty$. Therefore, since $\rho_k \to \infty$, there exists $k_1 \in \mathbb{N}$ such that $\|h_1(x_k)\|_\infty \leq \tau\|h_1(x_{k-1})\|_\infty$ for all $k \geq k_1$. So, $\rho_{k+1} = \rho_k$ for all $k \geq k_1$. Thus, $\{\rho_k\}$ is bounded. ∎

5.2. General constraints. In this subsection we address the general problem (2.1). As in the case of equality constraints, we adopt the following definition for the safeguarded Lagrange multipliers in Algorithm 3.1.

DEFINITION. *For all* $k \in \mathbb{N}$, $i = 1, \ldots, m_1$, $j = 1, \ldots, p_1$, $[\bar{\lambda}_{k+1}]_i$ *will be the projection of* $[\lambda_{k+1}]_i$ *on the interval* $[[\bar{\lambda}_{\min}]_i, [\bar{\lambda}_{\max}]_i]$, *and* $[\bar{\mu}_{k+1}]_j$ *will be the projection of* $[\mu_{k+1}]_j$ *on* $[0, [\bar{\mu}_{\max}]_j]$.

The technique for proving boundedness of the penalty parameter consists of reducing (2.1) to a problem with (only) equality constraints. The equality constraints of the new problem will be the active constraints at the limit point x_*. After this reduction, the boundedness result is deduced from Theorem 5.4. The sufficient conditions are listed below.

Assumption 7. The sequence $\{x_k\}$ is generated by the application of Algorithm 3.1 to problem (2.1) and $\lim_{k\to\infty} x_k = x_*$.

Assumption 8. The point x_* is feasible ($h_1(x_*) = 0$, $h_2(x_*) = 0$, $g_1(x_*) \leq 0$, and $g_2(x_*) \leq 0$.)

Assumption 9. The gradients $\{\nabla[h_1(x_*)]_i\}_{i=1}^{m_1}$, $\{\nabla[g_1(x_*)]_i\}_{[g_1(x_*)]_i=0}$, $\{\nabla[h_2(x_*)]_i\}_{i=1}^{m_2}$, $\{\nabla[g_2(x_*)]_i\}_{[g_2(x_*)]_i=0}$ are linearly independent. (LICQ holds at x_*.)

Assumption 10. The functions f, h_1, g_1, h_2, and g_2 admit continuous second derivatives in a neighborhood of x_*.

Assumption 11. Define the tangent subspace T as the set of all $z \in \mathbb{R}^n$ such that $\nabla h_1(x_*)^T z = \nabla h_2(x_*)^T z = 0$, $\langle \nabla[g_1(x_*)]_i, z\rangle = 0$ for all i such that $[g_1(x_*)]_i = 0$ and

$\langle \nabla [g_2(x_*)]_i, z\rangle = 0$ for all i such that $[g_2(x_*)]_i = 0$. Then, for all $z \in T, z \neq 0$,

$$\left\langle z, \left[\nabla^2 f(x_*) + \sum_{i=1}^{m_1} [\lambda_*]_i \nabla^2 [h_1(x_*)]_i + \sum_{i=1}^{p_1} [\mu_*]_i \nabla^2 [g_1(x_*)]_i \right.\right.$$
$$\left.\left. + \sum_{i=1}^{m_2} [v_*]_i \nabla^2 [h_2(x_*)]_i + \sum_{i=1}^{p_2} [u_*]_i \nabla^2 [g_2(x_*)]_i \right] z \right\rangle > 0.$$

Assumption 12. For all $i = 1, \dots, m_1, j = 1, \dots, p_1$, $[\lambda_*]_i \in ([\bar{\lambda}_{\min}]_i, [\bar{\lambda}_{\max}]_i)$, $[\mu_*]_j \in [0, [\bar{\mu}_{\max}]_j)$.

Assumption 13. For all i such that $[g_1(x_*)]_i = 0$, we have $[\mu_*]_i > 0$.

Observe that Assumption 13 imposes strict complementarity related only to the constraints in the upper-level set. In the lower-level set it is admissible that $[g_2(x_*)]_i = [u_*]_i = 0$. Observe, too, that Assumption 11 is weaker than the usual second-order sufficiency assumption, since the subspace T is orthogonal to the gradients of *all* active constraints, and no exception is made with respect to active constraints with null multiplier $[u_*]_i$. In fact, Assumption 11 is not a second-order sufficiency assumption for local minimizers. It holds for the problem of minimizing $x_1 x_2$ subject to $x_2 - x_1 \leq 0$ at $(0,0)$, although $(0,0)$ is not a local minimizer of this problem.

THEOREM 5.5. *Suppose that Assumptions* 7–13 *are satisfied. In addition, assume that there exists a sequence* $\eta_k \to 0$ *such that* $\varepsilon_k \leq \eta_k \max\{\|h_1(x_k)\|_\infty, \|\sigma_k\|_\infty\}$ *for all* $k \in \mathbb{N}$. *Then, the sequence of penalty parameters* $\{\rho_k\}$ *is bounded.*

Proof. Without loss of generality, assume that $[g_1(x_*)]_i = 0$ if $i \leq q_1$, $[g_1(x_*)]_i < 0$ if $i > q_1$, $[g_2(x_*)]_i = 0$ if $i \leq q_2$, and $[g_2(x_*)]_i < 0$ if $i > q_2$. Consider the auxiliary problem:

$$\text{Minimize } f(x) \text{ subject to } H_1(x) = 0, H_2(x) = 0, \tag{5.12}$$

where

$$H_1(x) = \begin{pmatrix} h_1(x) \\ [g_1(x)]_1 \\ \vdots \\ [g_1(x)]_{q_1} \end{pmatrix}, \quad H_2(x) = \begin{pmatrix} h_2(x) \\ [g_2(x)]_1 \\ \vdots \\ [g_2(x)]_{q_2} \end{pmatrix}.$$

By Assumptions 7–11, x_* satisfies the Assumptions 2–5 (with H_1, H_2 replacing h_1, h_2). Moreover, by Assumption 8, the multipliers associated to (2.1) are the Lagrange multipliers associated to (5.12).

As in the proof of (4.10) (the first part of the proof of Theorem 4.2), we have that, for k large enough, $[[g_1(x_*)]_i < 0 \Rightarrow [\mu_{k+1}]_i = 0]$ and $[[g_2(x_*)]_i < 0 \Rightarrow [u_k]_i = 0]$.

Then, by (3.1), (3.5), and (3.7),

$$\left\| \nabla f(x_k) + \sum_{i=1}^{m_1} [\lambda_{k+1}]_i \nabla [h_1(x_k)]_i + \sum_{i=1}^{q_1} [\mu_{k+1}]_i \nabla [g_1(x_k)]_i \right.$$
$$\left. + \sum_{i=1}^{m_2} [v_k]_i \nabla [h_2(x_k)]_i + \sum_{i=1}^{q_2} [u_k]_i \nabla [g_2(x_k)]_i \right\| \leq \varepsilon_k$$

for k large enough.

By Assumption 9, taking appropriate limits in the inequality above, we obtain that $\lim_{k\to\infty} \lambda_k = \lambda_*$ and $\lim_{k\to\infty} \mu_k = \mu_*$.

In particular, since $[\mu_*]_i > 0$ for all $i \leq q_1$,

$$[\mu_k]_i > 0 \tag{5.13}$$

for k large enough.

Since $\lambda_* \in (\bar{\lambda}_{\min}, \bar{\lambda}_{\max})^{m_1}$ and $[\mu_*]_i < [\bar{\mu}_{\max}]_i$, we have that $[\bar{\mu}_k]_i = [\mu_k]_i$, $i = 1, \ldots, q_1$, and $[\bar{\lambda}_k]_i = [\lambda_k]_i$, $i = 1, \ldots, m_1$, for k large enough.

Let us show now that the updating formula (3.7) for $[\mu_{k+1}]_i$, provided by Algorithm 3.1, coincides with the updating formula (3.5) for the corresponding multiplier in the application of the algorithm to the auxiliary problem (5.12).

In fact, by (3.7) and $[\bar{\mu}_k]_i = [\mu_k]_i$, we have that, for k large enough, $[\mu_{k+1}]_i = \max\{0, [\mu_k]_i + \rho_k [g_1(x_k)]_i\}$. But, by (5.13), $[\mu_{k+1}]_i = [\mu_k]_i + \rho_k [g_1(x_k)]_i, i = 1, \ldots, q_1$, for k large enough.

In terms of the auxiliary problem (5.12) this means that $[\mu_{k+1}]_i = [\mu_k]_i + \rho_k [H_1(x_k)]_i, i = 1, \ldots, q_1$, as we wanted to prove.

Now, let us analyze the meaning of $[\sigma_k]_i$. By (3.7), we have $[\sigma_k]_i = \max\{[g_1(x_k)]_i, -[\bar{\mu}_k]_i/\rho_k\}$ for all $i = 1, \ldots, p_1$. If $i > q_1$, since $[g_1(x_*)]_i < 0$, $[g_1]_i$ is continuous, and $[\bar{\mu}_k]_i = 0$, we have that $[\sigma_k]_i = 0$ for k large enough. Now, suppose that $i \leq q_1$. If $[g_1(x_k)]_i < -\frac{[\bar{\mu}_k]_i}{\rho_k}$, then, by (3.7), we would have $[\mu_{k+1}]_i = 0$. This would contradict (5.13). Therefore, $[g_1(x_k)]_i \geq -\frac{[\bar{\mu}_k]_i}{\rho_k}$ for k large enough, and we have that $[\sigma_k]_i = [g_1(x_k)]_i$. Thus, for k large enough,

$$H_1(x_k) = \begin{pmatrix} h_1(x_k) \\ \sigma_k \end{pmatrix}. \tag{5.14}$$

Therefore, the test for updating the penalty parameter in the application of Algorithm 3.1 to (5.12) coincides with the updating test in the application of the algorithm to (2.1). Moreover, formula (5.14) also implies that the condition $\varepsilon_k \leq \eta_k \max\{\|\sigma_k\|_\infty, \|h_1(x_k)\|_\infty\}$ is equivalent to the hypothesis $\varepsilon_k \leq \eta_k \|H_1(x_k)\|_\infty$ assumed in Theorem 5.4.

This completes the proof that the sequence $\{x_k\}$ may be thought of as being generated by the application of Algorithm 3.1 to (5.12). We proved that the associated approximate multipliers and the penalty parameters updating rule also coincide. Therefore, by Theorem 5.4, the sequence of penalty parameters is bounded, as we wanted to prove. □

Remark. The results of this section provide a theoretical answer to the following practical question: What happens if the box chosen for the safeguarded multiplier estimates is too small? The answer is that the box should be large enough to contain the "true" Lagrange multipliers. If it is not, the global convergence properties remain, but, very likely, the sequence of penalty parameters will be unbounded, leading to hard subproblems and possible numerical instability. In other words, if the box is excessively small, the algorithm tends to behave as an external penalty method. This is exactly what is observed in practice.

6. Numerical experiments. For solving unconstrained and bound-constrained subproblems we use GENCAN [9] with second derivatives and a CG-preconditioner [10]. Algorithm 3.1 with GENCAN will be called ALGENCAN. For solving the convex-constrained subproblems that appear in the large location problems, we use the spectral projected gradient method (SPG) [11, 12, 13]. The resulting augmented Lagrangian algorithm is called ALSPG. In general, it would be interesting to apply ALSPG to any problem such that the selected lower-level constraints define a convex

set for which it is easy (cheap) to compute the projection of an arbitrary point. The codes are free for download at http://www.ime.usp.br/~egbirgin/tango/. They are written in FORTRAN 77 (double precision). Interfaces of ALGENCAN with AMPL, CUTEr, C/C++, MATLAB, Octave, Python, and R (language and environment for statistical computing) are available.

For the practical implementation of Algorithm 3.1, we set $\tau = 0.5$, $\gamma = 10$, $\bar{\lambda}_{\min} = -10^{20}$, $\bar{\mu}_{\max} = \bar{\lambda}_{\max} = 10^{20}$, $\varepsilon_k = 10^{-4}$ for all k, $\bar{\lambda}_1 = 0, \bar{\mu}_1 = 0$, and $\rho_1 = \max\{10^{-6}, \min\{10, \frac{2|f(x_0)|}{\|h_1(x_0)\|^2 + \|g_1(x_0)_+\|^2}\}\}$. As stopping criterion we used $\max(\|h_1(x_k)\|_\infty, \|\sigma_k\|_\infty) \leq 10^{-4}$. The condition $\|\sigma_k\|_\infty \leq 10^{-4}$ guarantees that, for all $i = 1, \ldots, p_1$, $g_i(x_k) \leq 10^{-4}$ and that $[\mu_{k+1}]_i = 0$ whenever $g_i(x_k) < -10^{-4}$. This means that, approximately, feasibility and complementarity hold at the final point. Dual feasibility with tolerance 10^{-4} is guaranteed by (3.1) and the choice of ε_k. All the experiments were run on a 3.2 GHz Intel(R) Pentium(R) with four processors, 1Gb of RAM, and Linux Operating System. Compiler option "-O" was adopted.

6.1. Testing the theory. In Discrete Mathematics, experiments should reproduce exactly what theory predicts. In the continuous world, however, the situation changes because the mathematical model that we use for proving theorems is not exactly isomorphic to the one where computations take place. Therefore, it is always interesting to interpret, in finite precision calculations, the continuous theoretical results and to verify to what extent they are fulfilled.

Some practical results presented below may be explained in terms of a simple theoretical result that was tangentially mentioned in the introduction: If, at Step 2 of Algorithm 3.1, one computes a global minimizer of the subproblem and the problem (2.1) is feasible, then every limit point is a global minimizer of (2.1). This property may be easily proved using boundedness of the safeguarded Lagrange multipliers by means of external penalty arguments. Now, algorithms designed to solve reasonably simple subproblems usually include practical procedures that actively seek function decrease, beyond the necessity of finding stationary points. For example, efficient line-search procedures in unconstrained minimization and box-constrained minimization usually employ aggressive extrapolation steps [9], although simple backtracking is enough to prove convergence to stationary points. In other words, from good subproblem solvers one expects much more than convergence to stationary points. For this reason, we conjecture that augmented Lagrangian algorithms like ALGENCAN tend to converge to global minimizers more often than SQP-like methods. In any case, these arguments support the necessity of developing global-oriented subproblem solvers.

Experiments in this subsection were made using the AMPL interfaces of ALGENCAN (considering all the constraints as upper-level constraints) and IPOPT. Presolve AMPL option was disabled to solve the problems exactly as they are. The ALGENCAN parameters and stopping criteria were the ones stated at the beginning of this section. For IPOPT we used all its default parameters (including the ones related to stopping criteria). The random generation of initial points was made using the function `Uniform01()` provided by AMPL. When generating several random initial points, the seed used to generate the ith random initial point was set to i.

Example 1. Convergence to KKT points that do not satisfy MFCQ.

$$\begin{aligned} \text{Minimize} \quad & x_1 \\ \text{subject to} \quad & x_1^2 + x_2^2 \leq 1, \\ & x_1^2 + x_2^2 \geq 1. \end{aligned}$$

The global solution is $(-1, 0)$ and no feasible point satisfies the MFCQ, although all feasible points satisfy CPLD. Starting with 100 random points in $[-10, 10]^2$, ALGENCAN converged to the global solution in all the cases. Starting from $(5, 5)$ convergence occurred using 14 outer iterations. The final penalty parameter was 4.1649E-01 (the initial one was 4.1649E-03), and the final multipliers were 4.9998E-01 and 0.0000E+00. IPOPT also found the global solution in all the cases and used 25 iterations when starting from $(5, 5)$.

Example 2. Convergence to a non-KKT point.

$$\begin{aligned} \text{Minimize} \quad & x \\ \text{subject to} \quad & x^2 = 0, \\ & x^3 = 0, \\ & x^4 = 0. \end{aligned}$$

Here the gradients of the constraints are linearly dependent for all $x \in \mathbb{R}$. In spite of this, the only point that satisfies Theorem 4.1 is $x = 0$. Starting with 100 random points in $[-10, 10]$, ALGENCAN converged to the global solution in all the cases. Starting with $x = 5$ convergence occurred using 20 outer iterations. The final penalty parameter was 2.4578E+05 (the initial one was 2.4578E-05), and the final multipliers were 5.2855E+01, -2.0317E+00, and 4.6041E-01. IPOPT was not able to solve the problem in its original formulation because the "Number of degrees of freedom is NIND $= -2$." We modified the problem in the following way:

$$\begin{aligned} \text{Minimize} \quad & x_1 + x_2 + x_3 \\ \text{subject to} \quad & x_1^2 = 0, \\ & x_1^3 = 0, \\ & x_1^4 = 0, \\ & x_i \geq 0, \quad i = 1, 2, 3, \end{aligned}$$

and, after 16 iterations, IPOPT stopped near $x = (0, +\infty, +\infty)$ saying the "Iterates become very large (diverging?)."

Example 3. Infeasible stationary points [18, 34].

$$\begin{aligned} \text{Minimize} \quad & 100(x_2 - x_1^2)^2 + (x_1 - 1)^2 \\ \text{subject to} \quad & x_1 - x_2^2 \leq 0, \\ & x_2 - x_1^2 \leq 0, \\ & -0.5 \leq x_1 \leq 0.5, \\ & x_2 \leq 1. \end{aligned}$$

This problem has a global KKT solution at $x = (0, 0)$ and a stationary infeasible point at $x = (0.5, \sqrt{0.5})$. Starting with 100 random points in $[-10, 10]^2$, ALGENCAN converged to the global solution in all the cases. Starting with $x = (5, 5)$ convergence occurred using 6 outer iterations. The final penalty parameter was 1.0000E+01 (the initial one was 1.0000E+00), and the final multipliers were 1.9998E+00 and 3.3390E-03. IPOPT found the global solution starting from 84 out of the 100 random initial points. In the other 16 cases IPOPT stopped at $x = (0.5, \sqrt{0.5})$ saying "Convergence

to stationary point for infeasibility" (this was also the case when starting from $x = (5, 5)$).

Example 4. Difficult-for-barrier [15, 18, 47].

$$\begin{aligned} \text{Minimize} \quad & x_1 \\ \text{subject to} \quad & x_1^2 - x_2 + a = 0, \\ & x_1 - x_3 - b = 0, \\ & x_2 \geq 0,\ x_3 \geq 0. \end{aligned}$$

In [18] we read "This test example is from [47] and [15]. Although it is well-posed, many barrier-SQP methods ('Type-I Algorithms' in [47]) fail to obtain feasibility for a range of infeasible starting points."

We ran two instances of this problem, varying the values of parameters a and b and the initial point x_0 as suggested in [18]. When $(a, b) = (1, 1)$ and $x_0 = (-3, 1, 1)$ ALGENCAN converged to the solution $\bar{x} = (1, 2, 0)$ using 2 outer iterations. The final penalty parameter was 5.6604E-01 (the initial one also was 5.6604E-01), and the final multipliers were 6.6523E-10 and -1.0000E+00. IPOPT also found the same solution using 20 iterations. When $(a, b) = (-1, 0.5)$ and $x_0 = (-2, 1, 1)$ ALGENCAN converged to the solution $\tilde{x} = (1, 0, 0.5)$ using 5 outer iterations. The final penalty parameter was 2.4615E+00 (the initial one also was 2.4615E+00), and the final multipliers were -5.0001E-01 and -1.3664E-16. On the other hand, IPOPT stopped declaring convergence to a stationary point for the infeasibility.

Example 5. Preference for global minimizers.

$$\begin{aligned} \text{Minimize} \quad & \sum_{i=1}^{n} x_i \\ \text{subject to} \quad & x_i^2 = 1, \quad i = 1, \ldots, n. \end{aligned}$$

Solution: $x_* = (-1, \ldots, -1)$, $f(x_*) = -n$. We set $n = 100$ and ran ALGENCAN and IPOPT starting from 100 random initial points in $[-100, 100]^n$. ALGENCAN converged to the global solution in all the cases while IPOPT never found the global solution. When starting from the first random point, ALGENCAN converged using 4 outer iterations. The final penalty parameter was 5.0882E+00 (the initial one was 5.0882E-01), and the final multipliers were all equal to 4.9999E-01.

6.2. Location problems. Here we will consider a variant of the family of *location* problems introduced in [12]. In the original problem, given a set of n_p disjoint polygons $P_1, P_2, \ldots, P_{n_p}$ in $\mathbb{R}^2$ one wishes to find the point $z^1 \in P_1$ that minimizes the sum of the distances to the other polygons. Therefore, the original problem formulation is

$$\min_{z^i,\, i=1,\ldots,n_p} \quad \frac{1}{n_p - 1} \sum_{i=2}^{n_p} \|z^i - z^1\|_2 \quad \text{subject to} \quad z^i \in P_i,\ i = 1, \ldots, n_p.$$

In the variant considered in the present work, we have, in addition to the n_p polygons, n_c circles. Moreover, there is an ellipse which has a nonempty intersection with P_1 and such that z_1 must be inside the ellipse and $z_i, i = 2, \ldots, n_p + n_c$, must be outside. Therefore, the problem considered in this work is

$$\min_{z^i,\, i=1,\ldots,n_p+n_c} \quad \frac{1}{n_c + n_p - 1} \left[\sum_{i=2}^{n_p} \|z^i - z^1\|_2 + \sum_{i=1}^{n_c} \|z^{n_p+i} - z^1\|_2 \right]$$

TABLE 1

Location problems and their main features. The problem generation is based on a grid. The number of city-circles (n_c) and city-polygons (n_p) depend on the number of points in the grid, the probability of having a city in a grid point (procit), and the probability of a city to be a polygon (propol) or a circle ($1 - propol$). The number of vertices of a city-polygon is a random number and the total number of vertices of all the city-polygons together is totnvs. Finally, the number of variables of the problem is $n = 2(n_c + n_p)$, the number of upper-level inequality constraints is $p_1 = n_c + n_p$, and the number of lower-level inequality constraints is $p_2 = n_c + totnvs$. The total number of constraints is $p_1 + p_2$. The central rectangle is considered here a "special" city-polygon. The lower-level constraints correspond to the fact that each point must be inside a city and the upper-level constraints come from the fact that the central point must be inside the ellipse and all the others must be outside.

Problem	n_c	n_p	$totnvs$	n	p_1	p_2
1	28	98	295	252	126	323
2	33	108	432	282	141	465
3	33	108	539	282	141	572
4	33	109	652	284	142	685
5	35	118	823	306	153	858
6	35	118	940	306	153	975
7	35	118	1,057	306	153	1,092
8	35	118	1,174	306	153	1,209
9	35	118	1,291	306	153	1,326
10	35	118	1,408	306	153	1,443
11	35	118	1,525	306	153	1,560
12	35	118	1,642	306	153	1,677
13	35	118	1,759	306	153	1,794
14	35	118	1,876	306	153	1,911
15	35	118	1,993	306	153	2,028
16	35	118	2,110	306	153	2,145
17	35	118	2,227	306	153	2,262
18	35	118	2,344	306	153	2,379
19	3,029	4,995	62,301	16,048	8,024	65,330
20	4,342	7,271	91,041	23,226	11,613	95,383
21	6,346	10,715	133,986	34,122	17,061	140,332
22	13,327	22,230	278,195	71,114	35,557	291,522
23	19,808	33,433	417,846	106,482	53,241	437,654
24	29,812	50,236	627,548	160,096	80,048	657,360
25	26,318	43,970	549,900	140,576	70,288	576,218
26	39,296	66,054	825,907	210,700	105,350	865,203
27	58,738	99,383	1,241,823	316,242	158,121	1,300,561
28	65,659	109,099	1,363,857	349,516	174,758	1,429,516
29	98,004	164,209	2,052,283	524,426	262,213	2,150,287
30	147,492	245,948	3,072,630	786,880	393,440	3,220,122
31	131,067	218,459	2,730,798	699,052	349,526	2,861,865
32	195,801	327,499	4,094,827	1,046,600	523,300	4,290,628
33	294,327	490,515	6,129,119	1,569,684	784,842	6,423,446
34	261,319	435,414	5,442,424	1,393,466	696,733	5,703,743
35	390,670	654,163	8,177,200	2,089,666	1,044,833	8,567,870
36	588,251	979,553	12,244,855	3,135,608	1,567,804	12,833,106

$$\begin{aligned} \text{subject to} \quad & g(z^1) \leq 0, \\ & g(z^i) \geq 0, \quad i = 2, \ldots, n_p + n_c, \\ & z^i \in P_i, \quad i = 1, \ldots, n_p, \\ & z^{n_p+i} \in C_i, \quad i = 1, \ldots, n_c, \end{aligned}$$

where $g(x) = (x_1/a)^2 + (x_2/b)^2 - c$, and $a, b, c \in \mathbb{R}$ are positive constants. Observe that the objective function is differentiable in a large open neighborhood of the feasible region. To solve this family of problems, we will consider $g(z^1) \leq 0$ and $g(z^i) \geq$

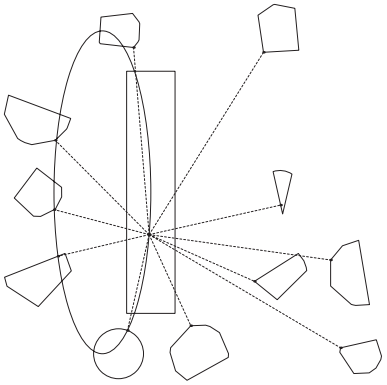


FIG. 1. *Twelve-sets very small location problem.*

$0, i = 2, \ldots, n_p + n_c$, as upper-level constraints, and $z^i \in P_i, i = 1, \ldots, n_p$, and $z^{n_p+i} \in C_i, i = 1, \ldots, n_c$, as lower-level constraints. In this way the subproblems can be efficiently solved by the SPG [11, 12] as suggested by the experiments in [12].

We generated 36 problems of this class, varying n_c and n_p and choosing randomly the location of the circles and polygons and the number of vertices of each polygon. Details of the generation, including the way in which we guarantee empty intersections (in order to have differentiability everywhere), may be found in [12] and its related code (also available at http://www.ime.usp.br/~egbirgin/tango/), where the original problem was introduced. Moreover, details of the present variant of the problem can be found within its fully commented FORTRAN 77 code (also available at http://www.ime.usp.br/~egbirgin/tango/). In Table 1 we display the main characteristics of each problem (number of circles, number of polygons, total number of vertices of the polygons, dimension of the problem, and number of lower-level and upper-level constraints). Figure 1 shows the solution of a very small twelve-sets problem that has 24 variables, 81 lower-level constraints, and 12 upper-level constraints.

The 36 problems are divided into two sets of 18 problems: small and large problems. We first solved the small problems with ALGENCAN (considering all the constraints as upper-level constraints) and ALSPG. Both methods use the FORTRAN 77 formulation of the problem (ALSPG needs an additional subroutine to compute the projection of an arbitrary point onto the convex set given by the lower-level constraints). In Table 2 we compare the performance of both methods for solving this problem. Both methods obtain feasible points and arrive at the same solutions. Due to the performance of ALSPG, we also used it to solve the set of large problems. Table 3 shows its performance. A comparison against IPOPT was made, and, while IPOPT was able to find equivalent solutions for the smaller problems, it was unable to handle the larger problems due to memory requirements.

7. Final remarks. In the last few years many sophisticated algorithms for nonlinear programming have been published. They usually involve combinations of interior-point techniques, SQP, trust regions, restoration, nonmonotone strategies, and advanced sparse linear algebra procedures. See, for example, [17, 28, 30, 31, 32, 37] and the extensive reference lists of these papers. Moreover, methods for solving efficiently specific problems or for dealing with special constraints are often introduced. Many times, a particular algorithm is extremely efficient for dealing with problems of a given type but fails (or cannot be applied) when constraints of a different class are

TABLE 2
Performance of ALGENCAN *and* ALSPG *in the set of small location problems.*

Problem	ALGENCAN					ALSPG					f
	OuIt	InIt	Fcnt	Gcnt	Time	OuIt	InIt	Fcnt	Gcnt	Time	
1	7	127	1309	142	1.21	3	394	633	397	0.10	1.7564E+01
2	6	168	1921	181	1.15	3	614	913	617	0.16	1.7488E+01
3	6	150	1818	163	1.02	3	736	1127	739	0.21	1.7466E+01
4	9	135	972	154	0.63	3	610	943	613	0.18	1.7451E+01
5	5	213	2594	224	1.16	3	489	743	492	0.15	1.7984E+01
6	5	198	2410	209	1.21	3	376	547	379	0.12	1.7979E+01
7	5	167	1840	178	1.71	3	332	510	335	0.11	1.7975E+01
8	3	212	2548	219	1.34	3	310	444	313	0.11	1.7971E+01
9	3	237	3116	244	1.49	3	676	1064	679	0.25	1.7972E+01
10	3	212	2774	219	1.27	3	522	794	525	0.20	1.7969E+01
11	3	217	2932	224	1.46	3	471	720	474	0.19	1.7969E+01
12	3	208	2765	215	1.40	3	569	872	572	0.23	1.7968E+01
13	3	223	2942	230	1.44	3	597	926	600	0.25	1.7968E+01
14	3	272	3981	279	2.12	3	660	1082	663	0.29	1.7965E+01
15	3	278	3928	285	2.20	3	549	834	552	0.24	1.7965E+01
16	3	274	3731	281	2.52	3	565	880	568	0.26	1.7965E+01
17	3	257	3186	264	2.31	3	525	806	528	0.24	1.7963E+01
18	3	280	3866	287	2.39	3	678	1045	681	0.32	1.7963E+01

TABLE 3
Performance of ALSPG *on set of large location problems. The memory limitation (to generate and save the problems' statement) is the only inconvenience for* ALSPG *solving problems with higher dimension than problem* 36 *(approximately* 3×10^6 *variables,* 1.5×10^6 *upper-level inequality constraints, and* 1.2×10^7 *lower-level inequality constraints), since computer time is quite reasonable.*

Problem	ALSPG					f
	OuIt	InIt	Fcnt	Gcnt	Time	
19	8	212	308	220	3.46	4.5752E+02
20	8	107	186	115	2.75	5.6012E+02
21	9	75	149	84	3.05	6.8724E+02
22	7	80	132	87	5.17	4.6160E+02
23	7	71	125	78	7.16	5.6340E+02
24	8	53	106	61	8.72	6.9250E+02
25	8	55	124	63	8.00	4.6211E+02
26	7	63	127	70	12.56	5.6438E+02
27	9	80	155	89	19.84	6.9347E+02
28	8	67	138	75	22.24	4.6261E+02
29	7	54	107	61	27.36	5.6455E+02
30	9	95	179	104	51.31	6.9382E+02
31	7	59	111	66	39.12	4.6280E+02
32	7	66	120	73	63.35	5.6449E+02
33	9	51	113	60	85.65	6.9413E+02
34	7	58	110	65	79.38	4.6270E+02
35	7	50	104	57	107.27	5.6432E+02
36	10	56	133	66	190.59	6.9404E+02

incorporated. This situation is quite common in engineering applications. In the augmented Lagrangian framework additional constraints are naturally incorporated into the objective function of the subproblems, which therefore preserve their constraint structure. For this reason, we conjecture that the augmented Lagrangian approach (with general lower-level constraints) will continue to be used for many years.

This fact motivated us to improve and analyze augmented Lagrangian methods with arbitrary lower-level constraints. From the theoretical point of view our goal was to eliminate, as much as possible, restrictive constraint qualifications. With this in mind, we used, both in the feasibility proof and in the optimality proof, the CPLD condition. This condition [41] has been proved to be a constraint qualification in [4], where its relations with other constraint qualifications have been given.

We provided a family of examples (location problems) where the potential of the arbitrary lower-level approach is clearly evidenced. This example represents a typical situation in applications. A specific algorithm (SPG) is known to be very efficient for a class of problems but turns out to be impossible to apply when additional constraints are incorporated. However, the augmented Lagrangian approach is able to deal with the additional constraints, taking advantage of the efficiency of SPG for solving the subproblems. In this way, we were able to solve nonlinear programming problems with more than 3,000,000 variables and 14,000,000 constraints in less than five minutes of CPU time.

Open problems related to theory and implementation of practical augmented Lagrangian methods may be found in the expanded report [3].

Acknowledgments. We are indebted to Prof. A. R. Conn, whose comments on a first version of this paper guided a deep revision, and to an anonymous referee for many constructive remarks.

REFERENCES

[1] M. Argáez and R. A. Tapia, *On the global convergence of a modified augmented Lagrangian linesearch interior-point method for nonlinear programming*, J. Optim. Theory Appl., 114 (2002), pp. 1–25.

[2] R. Andreani, E. G. Birgin, J. M. Martínez, and M. L. Schuverdt, *Augmented Lagrangian methods under the constant positive linear dependence constraint qualification*, Math. Program., 111 (2008), pp. 5–32.

[3] R. Andreani, E. G. Birgin, J. M. Martínez, and M. L. Schuverdt, *On Augmented Lagrangian Methods with General Lower-Level Constraints*, Technical report MCDO-051015, Department of Applied Mathematics, UNICAMP, Brazil, 2005; available online from http://www.ime.usp.br/∼egbirgin/.

[4] R. Andreani, J. M. Martínez, and M. L. Schuverdt, *On the relation between the constant positive linear dependence condition and quasinormality constraint qualification*, J. Optim. Theory Appl., 125 (2005), pp. 473–485.

[5] S. Bakhtiari and A. L. Tits, *A simple primal-dual feasible interior-point method for nonlinear programming with monotone descent*, Comput. Optim. Appl., 25 (2003), pp. 17–38.

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

[7] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed., Athena Scientific, Belmont, MA, 1999.

[8] E. G. Birgin, R. Castillo, and J. M. Martínez, *Numerical comparison of Augmented Lagrangian algorithms for nonconvex problems*, Comput. Optim. Appl., 31 (2005), pp. 31–56.

[9] E. G. Birgin and J. M. Martínez, *Large-scale active-set box-constrained optimization method with spectral projected gradients*, Comput. Optim. Appl., 23 (2002), pp. 101–125.

[10] E. G. Birgin and J. M. Martínez, *Structured minimal-memory inexact quasi-Newton method and secant preconditioners for augmented Lagrangian optimization*, Comput. Optim. Appl., to appear.

[11] E. G. Birgin, J. M. Martínez, and M. Raydan, *Nonmonotone spectral projected gradient methods on convex sets*, SIAM J. Optim., 10 (2000), pp. 1196–1211.
[12] E. G. Birgin, J. M. Martínez, and M. Raydan, *Algorithm* 813 : *SPG—Software for convex-constrained optimization*, ACM Trans. Math. Software, 27 (2001), pp. 340–349.
[13] E. G. Birgin, J. M. Martínez, and M. Raydan, *Inexact spectral projected gradient methods on convex sets*, IMA J. Numer. Anal., 23 (2003), pp. 539–559.
[14] R. H. Byrd, J. Ch. Gilbert, and J. Nocedal, *A trust region method based on interior point techniques for nonlinear programming*, Math. Program., 89 (2000), pp. 149–185.
[15] R. H. Byrd, M. Marazzi, and J. Nocedal, *On the convergence of Newton iterations to nonstationary points*, Math. Program., 99 (2004), pp. 127–148.
[16] R. H. Byrd, J. Nocedal, and A. Waltz, *Feasible interior methods using slacks for nonlinear optimization*, Comput. Optim. Appl., 26 (2003), pp. 35–61.
[17] R. H. Byrd, N. I. M. Gould, J. Nocedal, and R. A. Waltz, *An algorithm for nonlinear optimization using linear programming and equality constrained subproblems*, Math. Program., 100 (2004), pp. 27–48.
[18] L. Chen and D. Goldfarb, *Interior-point ℓ_2 penalty methods for nonlinear programming with strong global convergence properties*, Math. Program., 108 (2006), pp. 1–36.
[19] A. R. Conn, N. I. M. Gould, D. Orban, and Ph. L. Toint, *A primal-dual trust-region algorithm for nonconvex nonlinear programming*, Math. Program., 87 (2000), pp. 215–249.
[20] A. R. Conn, N. I. Gould, A. Sartenaer, and Ph. L. Toint, *Convergence properties of an augmented Lagrangian algorithm for optimization with a combination of general equality and linear constraints*, SIAM J. Optim., 6 (1996), pp. 674–703.
[21] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, *A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds*, SIAM J. Numer. Anal., 28 (1991), pp. 545–572.
[22] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, *Trust-Region Methods*, MPS/SIAM Ser. Optim. 1, SIAM, Philadelphia, 2000.
[23] M. A. Diniz-Ehrhardt, M. A. Gomes-Ruggiero, J. M. Martínez, and S. A. Santos, *Augmented Lagrangian algorithms based on the spectral projected gradient for solving nonlinear programming problems*, J. Optim. Theory Appl., 123 (2004), pp. 497–517.
[24] A. V. Fiacco and G. P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley & Sons, New York, 1968.
[25] R. Fletcher, *Practical Methods of Optimization*, Academic Press, London, 1987.
[26] R. Fletcher, N. I. M. Gould, S. Leyffer, Ph. L. Toint, and A. Wächter, *Global convergence of a trust-region SQP-filter algorithm for general nonlinear programming*, SIAM J. Optim., 13 (2002), pp. 635–659.
[27] A. Forsgren, P. E. Gill, and M. H. Wright, *Interior methods for nonlinear optimization*, SIAM Rev., 44 (2002), pp. 525–597.
[28] E. M. Gertz and P. E. Gill, *A primal-dual trust region algorithm for nonlinear optimization*, Math. Program., 100 (2004), pp. 49–94.
[29] P. E. Gill, W. Murray, and M. A. Saunders, *SNOPT: An SQP algorithm for large-scale constrained optimization*, SIAM Rev., 47 (2005), pp. 99–131.
[30] C. C. Gonzaga, E. Karas, and M. Vanti, *A globally convergent filter method for nonlinear programming*, SIAM J. Optim., 14 (2003), pp. 646–669.
[31] N. I. M. Gould, D. Orban, A. Sartenaer, and Ph. L. Toint, *Superlinear convergence of primal-dual interior point algorithms for nonlinear programming*, SIAM J. Optim., 11 (2001), pp. 974–1002.
[32] N. I. M. Gould, D. Orban, and Ph. L. Toint, *GALAHAD: A library of thread-safe Fortran* 90 *packages for large-scale nonlinear optimization*, ACM Trans. Math. Software, 29 (2003), pp. 353–372.
[33] M. R. Hestenes, *Multiplier and gradient methods*, J. Optim. Theory Appl., 4 (1969), pp. 303–320.
[34] W. Hock and K. Schittkowski, *Test Examples for Nonlinear Programming Codes*, Lecture Notes in Econom. and Math. Systems 187, Springer-Verlag, New York, 1981.
[35] X. Liu and J. Sun, *A robust primal-dual interior point algorithm for nonlinear programs*, SIAM J. Optim., 14 (2004), pp. 1163–1186.
[36] O. L. Mangasarian and S. Fromovitz, *The Fritz-John necessary optimality conditions in the presence of equality and inequality constraints*, J. Math. Anal. Appl., 17 (1967), pp. 37–47.
[37] J. M. Martínez, *Inexact restoration method with Lagrangian tangent decrease and new merit function for nonlinear programming*, J. Optim. Theory Appl., 111 (2001), pp. 39–58.
[38] J. M. Moguerza and F. J. Prieto, *An augmented Lagrangian interior-point method using directions of negative curvature*, Math. Program., 95 (2003), pp. 573–616.
[39] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, New York, 1999.

[40] M. J. D. Powell, *A method for nonlinear constraints in minimization problems*, in Optimization, R. Fletcher, ed., Academic Press, New York, 1969, pp. 283–298.
[41] L. Qi and Z. Wei, *On the constant positive linear dependence condition and its application to SQP methods*, SIAM J. Optim., 10 (2000), pp. 963–981.
[42] R. T. Rockafellar, *Augmented Lagrange multiplier functions and duality in nonconvex programming*, SIAM J. Control, 12 (1974), pp. 268–285.
[43] R. T. Rockafellar, *Lagrange multipliers and optimality*, SIAM Rev., 35 (1993), pp. 183–238.
[44] D. F. Shanno and R. J. Vanderbei, *Interior-point methods for nonconvex nonlinear programming: Orderings and high-order methods*, Math. Program., 87 (2000), pp. 303–316.
[45] P. Tseng, *Convergent infeasible interior-point trust-region methods for constrained minimization*, SIAM J. Optim., 13 (2002), pp. 432–469.
[46] M. Ulbrich, S. Ulbrich, and L. N. Vicente, *A globally convergent primal-dual interior-point filter method for nonlinear programming*, Math. Program., 100 (2004), pp. 379–410.
[47] A. Wächter and L. T. Biegler, *Failure of global convergence for a class of interior point methods for nonlinear programming*, Math. Program., 88 (2000), pp. 565–574.
[48] A. Wächter and L. T. Biegler, *On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming*, Math. Program., 106 (2006), pp. 25–57.
[49] R. A. Waltz, J. L. Morales, J. Nocedal, and D. Orban, *An interior algorithm for nonlinear optimization that combines line search and trust region steps*, Math. Program., 107 (2006), pp. 391–408.
[50] H. Yamashita and H. Yabe, *An interior point method with a primal-dual quadratic barrier penalty function for nonlinear optimization*, SIAM J. Optim., 14 (2003), pp. 479–499.

SIAM J. OPTIM.
Vol. 18, No. 4, pp. 1310–1325

ON HANDLING FREE VARIABLES IN INTERIOR-POINT METHODS FOR CONIC LINEAR OPTIMIZATION*

MIGUEL F. ANJOS† AND SAMUEL BURER‡

Abstract. We revisit a regularization technique of Mészáros for handling free variables within interior-point methods for conic linear optimization. We propose a simple computational strategy, supported by a global convergence analysis, for handling the regularization. Using test problems from benchmark suites and recent applications, we demonstrate that the modern code SDPT3 modified to incorporate the proposed regularization is able to achieve the same or significantly better accuracy over standard options of splitting variables, using a quadratic cone, and solving indefinite systems.

Key words. infeasible primal-dual path-following algorithm, semidefinite programming, equality constraints, free variables, regularization

AMS subject classifications. 90C51, 90C22, 90C05, 65K05

DOI. 10.1137/06066847X

1. Introduction.

Conic linear optimization, and in particular semidefinite optimization, has arisen since the early 1990s as an increasingly powerful and useful technique for tackling a variety of problems arising from both applications and theory. We refer the reader to the SDP webpage of Helmberg [8] as well as the books of de Klerk [5] and Wolkowicz, Saigal, and Vandenberghe [30] for thorough coverage of the theory and algorithms in this area, as well as of several application areas where researchers in conic linear optimization have made significant contributions.

Conic linear optimization refers to the class of optimization problems where a linear function of a variable x is optimized subject to linear constraints on the elements of x and the additional constraint that x lie in a symmetric self-dual cone. This includes linear programming (LP) problems as a special case, namely, when the cone is the nonnegative orthant. Since all of these cones can be described as a conic section of the cone of positive semidefinite matrices (in a polynomially bounded dimension, see [6]), attention has focused particularly on the development of algorithms for solving semidefinite linear optimization, commonly referred to as semidefinite programming (SDP). Beyond LP and SDP, a third specific cone that is useful in applications is the second-order (or Lorentz) cone, which gives rise to second-order cone programming (SOCP).

As a result, a variety of algorithms for solving LP, SOCP, and SDP problems, including polynomial-time infeasible path-following interior-point methods (IPMs), have been implemented and benchmarked (see, e.g., [18]), and several excellent solvers are available. Two of these solvers handle LP, SOCP, and SDP in a unified way, namely, SeDuMi [25] and SDPT3 [26].

Notwithstanding the substantial progress made in recent years, work continues on methods and software for conic linear optimization. One outstanding issue is that of handling free variables.

*Received by the editors August 27, 2006; accepted for publication (in revised form) August 3, 2007; published electronically November 7, 2007.
http://www.siam.org/journals/siopt/18-4/66847.html

†Department of Management Sciences, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada (anjos@stanfordalumni.org). Research partially supported by Discovery grant 312125 and RTI grant 314668 from the Natural Sciences and Engineering Research Council of Canada.

‡Department of Management Sciences, University of Iowa, Iowa City, IA 52242-1994 (samuel-burer@uiowa.edu). Research partially supported by NSF grants CCR-0203426 and CCF-0545514.

Handling free variables in conic linear optimization is an important modeling and algorithmic issue for IPMs. The issue arises from the fact that nearly all theories and algorithms for IPMs are based on the following standard-form primal and dual problems:

$$\min\left\{c^T x : Ax = b,\ x \in K\right\} \qquad \text{and} \qquad \max\left\{b^T y : A^T y + s = c,\ s \in K\right\},$$

where K is the symmetric self-dual cone (specifically, a direct product of linear, semidefinite, and second-order cones). However, in many applications, free variables naturally appear in the primal. Examples of such applications occur in quantum chemistry [36], polynomial optimization problems [21, 12], and combinatorial optimization problems [13, 2], among others.

Allowing free variables, the (nonstandard) primal problem is

$$\min\left\{c^T x + g^T z : Ax + Ez = b,\ x \in K\right\} \tag{1}$$

with the corresponding dual problem

$$\max\left\{b^T y : A^T y + s = c,\ E^T y = g,\ s \in K\right\}. \tag{2}$$

We denote by (n, p, m, n) the dimensions of the vectors (x, z, y, s), respectively, which determines the sizes of the data (A, b, c, E, g).

Theoretically, it is not so difficult to extend standard-form IPMs to handle (1)–(2). However, computation is not so easy. Each iteration of IPMs for the standard-form problem is based on solving a positive definite linear system, and accordingly most codes use high-quality, fast, sparsity-preserving implementations of the Cholesky factorization. When free variables are present, the corresponding system is still invertible but becomes indefinite, which makes it more difficult to solve in a quick, stable fashion. We emphasize that free variables are a computational issue and not a theoretical one.

Researchers have attempted various alternative ways to handle free variables computationally. Each method can be viewed as an attempt to enable the use of the Cholesky factorization.

Probably the simplest and most often suggested method is to split the variable z into a difference $z^+ - z^-$ of nonnegative vectors $z^+ \geq 0$ and $z^- \geq 0$ which transforms (1)–(2) into standard form. The effect in the dual is that the equality $E^T y = g$ is split into $E^T y \geq g$ and $E^T y \leq g$. In part because of its simplicity, this approach is often implemented as the default behavior for interior-point methods (e.g., in the LP code LIPSOL [34] that is the basis of Matlab's large-scale LP solver). For interior-point methods, this splitting of variables can be problematic because the primal optimal solution set becomes unbounded and the dual feasible set has no interior. Empirically, a typical behavior is that z^+ and z^- individually become unbounded, while their difference z stays bounded. For LP, Wright [31] asserts that, for methods which achieve superlinear convergence, the tendency of z^+ and z^- to grow large is mitigated in a satisfactory manner. Also, an alternative way to handle the splitting of variables in LP that leads to the solution of a symmetric quasi-definite system [27] is outlined in [28]. However, the recent results in [29, 9] suggest that the degeneracy caused by splitting free variables in SDP problems makes it difficult to solve the resulting SDPs stably and/or highly accurately.

A quick overview of other alternative methods for handling free variables is as follows:

- One can convert to standard form by eliminating z in the equations $Ax+Ez = b$. However, structural properties of (A, E) such as sparsity tend to be destroyed by such an approach. In the context of SDP, Kobayashi, Nakata, and Kojima [9] consider this approach and in particular make efforts to manage the loss of sparsity by eliminating z via different bases.
- One can convert to standard form by adding an auxiliary scalar variable z_0 and requiring that (z_0, z) be in a second-order cone. To the best of our knowledge, this was first suggested by Andersen [1], who reports good results and improved accuracy when solving SOCPs. This approach is also available as an option in the most recent release of SeDuMi, version 1.1 (see [23]).
- One can handle the free variables directly and regularize the indefinite system faced at each iteration, i.e., make the system symmetric quasi-definite by perturbing it in a controlled way. This approach was suggested by Mészáros [17] in the context of LP. Beyond permitting the use of the Cholesky factorization, another advantage of regularization is that the structure of (A, E) is not destroyed. On the other hand, the downside of this approach is that the solution of the system is also perturbed, and care must be taken to ensure that the global convergence of the method is not negatively affected.

Finally, we reiterate that one certainly still has the option to handle the free variables directly and solve the indefinite systems ((7) below). In fact, this is the default option of SDPT3 version 3.02, ostensibly because the authors of the software found this approach better than, say, splitting variables.

It seems safe to say that no consensus has been reached on how to handle free variables in all cases. Indeed, it is our opinion (and that of Kobayashi, Nakata, and Kojima [9]) that solving general conic linear optimization problems with free variables in a reliably stable and accurate manner remains a relevant research topic.

In this paper, we revisit the regularization method of Mészáros [17] and formalize a strategy for handling and updating the regularization so that global convergence is not affected. In contrast to the strategy suggested by Mészáros, which is based on iterative refinement and is somewhat ad hoc, our strategy is supported by a global convergence result. Using the code SDPT3, we illustrate the effectiveness of our regularization strategy on a diverse collection of problems. Our approach achieves the same or significantly better accuracy over the approaches of splitting variables, using a quadratic cone, and solving indefinite systems.

1.1. Some remarks. A few remarks are in order. First, our intention in this paper is not to claim that regularization is the best in all situations. Indeed, this is most likely *not* the case. Instead, we simply hope to establish that regularization, properly handled, is a viable alternative to other methods for handling free variables. (One consequence is that we have chosen not to compare with the method of eliminating z as in [9] in part because careful comparisons are given in [9] and because we prefer to maintain the structure of (A, E).)

Second, there are several different publicly available codes for LP, SOCP, and/or SDP on which we could test the regularization. Ultimately, we have chosen to test SDPT3 for several reasons: We wished to test both SOCP and SDP, which SDPT3 can handle; SDPT3's algorithm matches the algorithmic framework of our analysis very closely; and SDPT3's code is easily accessible, customizable, and verifiable in Matlab.

Finally, significant variation between different codes makes it unclear whether regularization would have the same effect within all codes as it does with SDPT3.

For example, tests that we have performed indicate that split variables perform quite well within SeDuMi. This is to be expected because, by design, the homogeneous self-dual embedding model used by SeDuMi has a bounded optimal solution set, and thus the iterates cannot diverge to infinity. Hence, SeDuMi will not suffer additional numerical instabilities from the splitting; in fact, split variables perform so well in SeDuMi that there does not appear to be much room for improvement in SeDuMi using regularization. In our opinion, these code-by-code differences make the discussion of free variables richer.

1.2. Structure of this paper. This paper is structured as follows. In section 2, we recall the basic framework of infeasible primal-dual path-following algorithms. In section 3, we summarize the key ideas behind a global convergence result of Kojima, Megiddo, and Mizuno [10], and in section 4, we recall the regularization approach originally proposed by Mészáros [17]. In section 5, we propose a specific methodology to update the regularization at each iteration and extend the analysis of Kojima, Megiddo, and Mizuno [10] to show that the resulting infeasible primal-dual path-following method is globally convergent. In section 6, we report computational results which show that the proposed regularization leads to an overall improvement in the performance of SDPT3 for instances with free variables. Finally, section 7 summarizes our findings and mentions some possible directions for future research.

2. The basic infeasible primal-dual path-following framework. In this section, we recall the basic framework of infeasible primal-dual path-following algorithms, which is implemented in nearly all interior-point codes for conic linear optimization. Even though standard texts treat the standard-form problem, we state the framework with respect to the problems (1)–(2). To recover the standard-form framework, one can simply take $p = 0$.

For simplicity, the framework (and indeed all of the results in the paper) are stated with K expressed as a linear cone; i.e., the SOCP and SDP cases are not explicitly handled. By now, it is well known that all standard convergence results for LP can be extended to SOCP and SDP. With this in mind and in hopes of keeping this paper as clean and accessible as possible, we choose to state everything in terms of LP.

Without loss of generality, we make the standard assumptions that A has full-row rank and E has full-column rank. We also assume that both (1)–(2) are interior feasible so that strong duality holds. Strong duality occurs when both primal and dual attain their optimal values with no gap; i.e., there exists a primal-dual feasible point (x, z, y, s) such that $\mu(x, s) = 0$, where $\mu(x, s) := x^T s/n$ is the (scaled) duality gap.

A consequence of these assumptions is that, for all $\nu > 0$, the system

$$Ax + Ez = b, \tag{3a}$$

$$A^T y + s = c, \tag{3b}$$

$$E^T y = g, \tag{3c}$$

$$XS\,e = \nu\,e, \tag{3d}$$

$$(X, S) \in K^0 \times K^0 \tag{3e}$$

has a unique solution, which we write as $(x_\nu, z_\nu, y_\nu, s_\nu)$. (We adopt common notation in the field of IPMs, so that $X := \mathrm{Diag}(x)$, $S := \mathrm{Diag}(s)$, $K^0 := \mathrm{int}(K)$, and e denotes the vector of all ones.) The set $\mathcal{C} := \{(x_\nu, z_\nu, y_\nu, s_\nu) : \nu > 0\}$ is called the *central path* and is a smooth trajectory that converges to the primal-dual optimal solution set as $\nu \to 0$ (note, for example, that $\mu(x_\nu, s_\nu) = \nu$).

Given an initial point (x^1, z^1, y^1, s^1)—which is not necessarily primal-dual feasible but does satisfy $(x^1, s^1) \in K^0 \times K^0$—the kth iteration of the path-following framework attempts to solve (3) for some $\nu_k \in (0, \mu(x^k, s^k))$ by taking a step via Newton's method. More specifically, the system

$$A\Delta x^k + E\Delta z^k = r_p^k, \tag{4a}$$

$$A^T\Delta y^k + \Delta s^k = r_{d_1}^k, \tag{4b}$$

$$E^T\Delta y^k = r_{d_2}^k, \tag{4c}$$

$$S^k\Delta x^k + X^k\Delta s^k = r_c^k, \tag{4d}$$

where

$$r_p^k := b - Ax^k - Ez^k, \tag{5a}$$

$$r_{d_1}^k := c - A^T y^k - s^k, \tag{5b}$$

$$r_{d_2}^k := g - E^T y^k, \tag{5c}$$

$$r_c^k := \nu_k\, e - X^k S^k e, \tag{5d}$$

is solved for $(\Delta x^k, \Delta z^k, \Delta y^k, \Delta s^k)$, and a step size $\alpha_k \in (0, 1]$ is selected such that

$$(x^{k+1}, z^{k+1}, y^{k+1}, s^{k+1}) := (x^k, z^k, y^k, s^k) + \alpha_k(\Delta x^k, \Delta z^k, \Delta y^k, \Delta s^k)$$

satisfies $\mu(x^{k+1}, s^{k+1}) < \mu(x^k, s^k)$ and $(x^{k+1}, s^{k+1}) \in K^0 \times K^0$. By construction,

$$(r_p^{k+1}, r_{d_1}^{k+1}, r_{d_2}^{k+1}) = (1 - \alpha_k)(r_p^k, r_{d_1}^k, r_{d_2}^k). \tag{6}$$

In other words, one can interpret a single iteration as decreasing the duality gap and decreasing infeasibility, while staying inside the cone. (An important technical issue is whether (4) is uniquely solvable. This is guaranteed by $(x^k, s^k) \in K^0 \times K^0$; see also below.)

Various implementations of the above basic framework are possible. For example, many implementations do not monitor the decrease of μ since, in practice, a decrease is typically observed all the way to the boundary of $K^0 \times K^0$. Most implementations also take different step sizes in the primal and dual spaces. Another popular variant is the predictor-corrector strategy of Mehrotra [16], which provides a highly effective scheme for choosing ν_k and for altering $(\Delta x^k, \Delta z^k, \Delta y^k, \Delta s^k)$ so that the central path is followed more closely.

The Newton system can be reduced to the following smaller system (where the superscript k is understood):

$$\begin{pmatrix} AXS^{-1}A^T & E \\ E^T & 0 \end{pmatrix} \begin{pmatrix} \Delta y \\ \Delta z \end{pmatrix} = \begin{pmatrix} AXS^{-1}r_{d_1} - AS^{-1}r_c + r_p \\ r_{d_2} \end{pmatrix}. \tag{7}$$

If $p > 0$ (i.e., if there are free variables), then this system is indefinite. On the other hand, if $p = 0$, then the system is positive definite and can be solved with the Cholesky factorization.

3. A basic global convergence analysis. As discussed in section 2, the infeasible primal-dual path-following framework reduces both the gap and the infeasibility in each iteration. Global convergence is achieved if the gap and infeasibility converge to zero. In this section, we recapitulate the first global convergence result, which

was given by Kojima, Megiddo, and Mizuno [10] (for the case of LP with no free variables). This will serve as the basis of the results in section 5. (Some comments on why we have chosen to present the approach of Kojima, Megiddo, and Mizuno are given below in section 3.2.)

We caution the reader that we do not present Kojima, Megiddo, and Mizuno's method verbatim, but instead we make some simplifying assumptions. The simplifications are for the sake of brevity; all of the fundamental content is retained. For example, Kojima, Megiddo, and Mizuno analyze the use of different primal and dual step sizes, whereas we analyze a common step size. Another simplification is the following assumption: The initial point (x^1, z^1, y^1, s^1) satisfies $(r_p^1, r_{d_1}^1) = (0, 0)$ so that $(r_p^k, r_{d_1}^k) = (0, 0)$ for all k. Our reasons for this assumption are as follows: The essentials of the global convergence result are clear with only one infeasible equation, and the assumption $r_{d_2}^k \neq 0$ is sufficient to develop the techniques of section 5. In accordance with this assumption, we will write $r^k := r_{d_2}^k$ to streamline notation. Also to make notation easier, we let $\mu_k := \mu(x^k, s^k)$.

Within the framework of section 2, convergence results typically require some restrictions on the iterates. Kojima, Megiddo, and Mizuno require that the iterates remain in a neighborhood of the central path having the following form, which is dependent on constants $\gamma \in (0, 1)$ and $\beta > 0$:

(8)
$$\mathcal{N}(\gamma, \beta) := \{(x, z, y, s) \in K^0 \times \Re^p \times \Re^m \times K^0 \ : \ XSe \geq \gamma\, \mu(x, s)\, e,\ \|r(y)\| \leq \beta\, \mu(x, s)\},$$

where $r(y) := g - E^T y$. In particular, γ and β should be chosen so that $(x^1, z^1, y^1, s^1) \in \mathcal{N}(\gamma, \beta)$. At times we will write $\mathcal{N} := \mathcal{N}(\gamma, \beta)$ for convenience. The neighborhood aids the convergence analysis by guaranteeing that the iterates do not get too close to the cone boundary and that infeasibility decreases at the same rate as the duality gap.

The precise algorithm is stated as Algorithm 1. Note that the algorithm depends on user-defined tolerances $\varepsilon > 0$ and $\omega > 0$ as well as a duality gap "dampening" factor $\sigma \in (0, 0.99)$. In addition, the algorithm also utilizes the following definitions for $\alpha \in [0, 1]$:

$$\begin{aligned}(x_\alpha, z_\alpha, y_\alpha, s_\alpha) &:= (x^k, z^k, y^k, s^k) + \alpha(\Delta x^k, \Delta z^k, \Delta y^k, \Delta s^k),\\ \mu_\alpha &:= \mu(x_\alpha, s_\alpha).\end{aligned}$$

The convergence result is stated next.

Algorithm 1. Infeasible Path-Following Algorithm.

Let $\varepsilon > 0$, $\omega > 0$, $\sigma \in (0, 0.99)$, and $(x^1, z^1, y^1, s^1) \in \mathcal{N}$ be given.

for $k = 1, 2, 3, \ldots$ **do**

If $\mu_k \leq \varepsilon$ or $\|(x^k, s^k)\|_1 \geq \omega$, then stop.

Set $\nu_k := \sigma\mu_k$ and solve (4) for $(\Delta x^k, \Delta z^k, \Delta y^k, \Delta s^k)$.

Set $(x^{k+1}, z^{k+1}, y^{k+1}, s^{k+1}) = (x_{\alpha_k}, z_{\alpha_k}, y_{\alpha_k}, s_{\alpha_k})$, where $\alpha_k \in (0, 1]$ is the largest step size such that the relations

$$(x_\alpha, z_\alpha, y_\alpha, s_\alpha) \in \mathcal{N}, \tag{9}$$

$$\mu_\alpha \leq (1 - 0.01\alpha)\mu_k \tag{10}$$

hold for every $\alpha \in [0, \alpha_k]$.

end for

THEOREM 3.1 (see [10]). *Let $\gamma \in (0,1)$, $\beta > 0$, and an initial point $(x^1, z^1, y^1, s^1) \in \mathcal{N}(\gamma, \beta)$ be given. Suppose, moreover, that positive tolerances ε and ω and a dampening factor $\sigma \in (0, 0.99)$ are specified. Then Algorithm 1 eventually generates an iterate $(x^k, z^k, y^k, s^k) \in \mathcal{N}(\gamma, \beta)$ such that $\mu_k \leq \varepsilon$ or $\|(x^k, s^k)\|_1 \geq \omega$. If the first case occurs, then $\|r^k\| \leq \beta\,\varepsilon$ as well.*

If the second case occurs, then Kojima, Megiddo, and Mizuno show that the infeasibility of (1)–(2) is implied over a wide region of the primal-dual ground space $K \times \Re^p \times \Re^m \times K$. Although this information is not a full infeasibility certificate, the intuition is that this is a strong indication of infeasibility, especially when ω is large.

The key to establishing Theorem 3.1 is to prove the existence of a positive constant α_* such that $\alpha_k \geq \alpha_*$ for all k generated by the algorithm. We state this lemma and prove the theorem; portions of the proof of the lemma, which are relevant to section 5, are given below in section 3.1.

LEMMA 3.2 (see [10]). *Suppose that there exists some constant $\eta > 0$ such that, for all k generated by the algorithm,*

$$|\,\Delta x_i^k \Delta s_i^k - \gamma\,(\Delta x^k)^T \Delta s^k/n\,| \leq \eta \quad \forall\ i = 1, \ldots, n, \tag{11a}$$

$$|\,(\Delta x^k)^T \Delta s^k/n\,| \leq \eta. \tag{11b}$$

Then $\alpha_k \geq \alpha_* > 0$, where

$$\alpha_* := \min\left\{1, \frac{(1-\gamma)\sigma\varepsilon}{\eta}, \frac{\sigma\varepsilon}{\eta}, \frac{(0.99-\sigma)\varepsilon}{\eta}\right\}. \tag{12}$$

Proof of Theorem 3.1. The proof is by contradiction. Assume that Algorithm 1 does not terminate. Then the entire infinite sequence $\{(x^k, z^k, y^k, s^k)\}$ lies in the compact set

$$\mathcal{N}^* := \{\ (x,z,y,s) \in \mathcal{N}\ :\ \mu(x,s) \geq \varepsilon,\ \|(x,s)\|_1 \leq \omega\ \}.$$

Combining this with the fact that the Newton direction is a continuous function of the iterates (since (4) is nonsingular for each k), any continuous function of the direction is uniformly bounded over all k. So the hypothesis of Lemma 3.2 holds, implying that $\alpha_k \geq \alpha_*$ for all k. Thus, by Algorithm 1, the duality gap is decreased by at least a multiplicative factor of $1 - 0.01\alpha_* < 1$ in each iteration, and so $\mu_k \to 0$, which contradicts the assumption that Algorithm 1 does not terminate. ∎

3.1. Proof of Lemma 3.2. Assume that (11) holds for all k generated by the algorithm, and note also that

$$\mu_k \geq \varepsilon \tag{13}$$

for the same k. We define $r_\alpha := g - E^T y_\alpha$ and recall that $r_\alpha = (1-\alpha)\,r$ by (6).

Using the definitions of $\mathcal{N}$ and Algorithm 1 together with the following three propositions, the proof of Lemma 3.2 is straightforward. We give only the proof of Proposition 3.4 because of its relevance for the results in section 5. The super- and subscripts k are dropped since the arguments below are irrespective of k.

PROPOSITION 3.3 (see [10]). *$X_\alpha S_\alpha\, e \geq \gamma\,\mu_\alpha\, e$ for all $\alpha \leq (1-\gamma)\,\sigma\,\varepsilon/\eta$.*

PROPOSITION 3.4 (see [10]). *$\|r_\alpha\| \leq \beta\mu_\alpha$ for all $\alpha \leq \sigma\varepsilon/\eta$.*

Proof. Recall the standard relation

$$\mu_\alpha = (1 - (1-\sigma)\alpha)\mu + \alpha^2 \Delta x^T \Delta s/n. \tag{14}$$

Thus, we have

$$\begin{aligned}\beta\mu_\alpha - \|r_\alpha\| &= \beta\left[(1-(1-\sigma)\alpha)\mu + \alpha^2\Delta x^T\Delta s/n\right] - (1-\alpha)\|r\| \\ &= (1-\alpha)(\beta\mu - \|r\|) + \beta\alpha\sigma\mu + \beta\alpha^2\Delta x^T\Delta s/n \\ &\geq \beta\alpha\sigma\mu - \beta\alpha^2\eta \geq \alpha\beta\left[\sigma\varepsilon - \alpha\eta\right],\end{aligned}$$

where the first equality follows from (14), the first inequality follows from $(x, z, y, s) \in \mathcal{N}$ and (11b), and the second inequality follows from (13). This proves the result. ∎

PROPOSITION 3.5 (see [10]). *$\mu_\alpha \leq (1 - 0.01\alpha)\mu$ for all $\alpha \leq (0.99 - \sigma)\varepsilon/\eta$.*

3.2. Other convergence results. Before proceeding, we discuss other known convergence results for the infeasible framework of section 2 and explain why we have chosen to analyze the result of Kojima, Megiddo, and Mizuno [10].

Following the above global convergence result of Kojima, Megiddo, and Mizuno and Zhang [32] strengthened the result by proving that an ε-approximate optimal solution is delivered within $O(n^2)$ iterations (if an optimal solution exists). In particular, Zhang resolved the ambiguity surrounding Kojima, Megiddo, and Mizuno's infeasibility "certificate" $\|(x, s)\|_1 > \omega$. For example, under the assumption of primal-dual interior feasibility, all iterates stay bounded. Later, Zhang [33] extended these ideas to the case of SDP.

Other implementations of the framework have also been analyzed. In particular, a few authors have studied variations of the original predictor-corrector strategy of Mehrotra [16]. First, Mehrotra himself suggests a proof of global convergence for his method by appealing to certain "fall back" search directions, which are different from his own predictor-corrector direction. Then Zhang and Zhang [35] prove polynomial convergence of a variation of Mehrotra's original method, which they suggest, but to our knowledge the Zhang–Zhang variant has not actually been implemented in practice. Finally, a third variant, which is used in most modern interior-point codes, is analyzed by Salahi, Peng, and Terlaky [24]. They make the simplifying assumption that all iterates are feasible and show by example that this variant can converge quite slowly on certain problems. By studying a suitable modification, they prove polynomial convergence. To our knowledge, no one has proved global or polynomial convergence of the infeasible version of this third variant (i.e., the one implemented in most codes).

A different line of research has analyzed the convergence of the framework when using inexact Newton directions, which are directions that satisfy (4) only approximately. Inexact directions arise, for example, when iterative methods are used to solve the system (7) to moderate accuracy. Depending on their precise form, inexact directions can lead to infeasible iterates, even if the algorithm is supplied with an initial feasible iterate. This is because the key relation (6) does not hold from iteration to iteration. Nevertheless, global and polynomial convergence results can be proved under suitable conditions on the degree of the inexactness of the direction (see [19, 11, 7] for LP and, more recently, [37] for SDP).

In section 5, we propose Algorithm 2, a variant of Algorithm 1 that is based on an inexact Newton direction arising from regularization. We also extend the analysis in this section to prove a global convergence result for Algorithm 2. Although we were unable to prove a polynomial convergence result for Algorithm 2, and although the aforementioned work on inexact Newton directions can be applied to obtain a provably polynomially convergent method with regularization, we deliberately choose to advocate Algorithm 2 for the following reasons:

- Our extension of the original result of Kojima, Megiddo, and Mizuno has the advantages that: (i) it allows for infeasible iterates and a straightforward analysis of our particular inexact direction; and (ii) the resulting regularization strategy is quite simple to implement and has an intuitive appeal.
- Our experiments in the direction of following the insights provided by the above research on inexact methods led us to the following conclusions: (i) the resulting regularization strategies are significantly more difficult to implement; and (ii) we implemented the approach of [7] and found by experimentation that it did not work as well as our proposed approach.

It should also be noted that we chose *not* to analyze a predictor-corrector variant because the theoretical basis for implemented predictor-corrector strategies is less well understood, especially with regards to infeasible and inexact aspects. Nonetheless, we did test our method successfully within such a strategy (more details are given in section 6).

Ultimately, we believe that our analysis allows us to identify the essence of a good regularization strategy. It should (i) be convergent, (ii) be easy to implement, and (iii) work well in practice.

4. The regularization approach of Mészáros. Mészáros [17] proposes to replace (4c) of the Newton system (4) with the following equation for a specified $\delta_k > 0$:

$$E^T \Delta y^k - \delta_k \, \Delta z^k = r_{d_2}^k. \tag{15}$$

Just as (4) can be reduced to (7), the regularized system of Mészáros can be reduced to

$$\begin{pmatrix} AXS^{-1}A^T & E \\ E^T & -\delta\, I \end{pmatrix} \begin{pmatrix} \Delta y \\ \Delta z \end{pmatrix} = \begin{pmatrix} AXS^{-1}r_{d_1} - AS^{-1}r_c + r_p \\ r_{d_2} \end{pmatrix}, \tag{16}$$

where the k subscript is understood. In contrast to (7), however, (16) can further be reduced to the positive definite system

$$\left(AXS^{-1}A^T + \delta^{-1}EE^T\right) \Delta y = AXS^{-1}r_{d_1} - AS^{-1}r_c + r_p + \delta^{-1}Er_{d_2}. \tag{17}$$

Hence, the Cholesky factorization can be employed to calculate the direction. The obvious downside is that the resulting direction is not the true Newton direction. It is shown in [17] that the difference between these two directions is $O(\delta)$.

Yet, a more important question in practice is the choice of δ_k throughout the course of the algorithm, since poor choices can certainly have a negative impact on convergence. Under restrictive assumptions, an adaptive heuristic for updating δ_k is proposed in [17], and iterative refinement is also suggested for improving the quality of the search direction at each iteration. Furthermore, Maros and Mészáros [15] investigate the choice of a constant δ_k throughout the algorithm. From our perspective, these suggestions are somewhat ad hoc and do not consider the effect of the regularization on the global convergence of the algorithm. We feel that the question of how to select a global strategy for updating δ_k was left open by Mészáros.

5. Global convergence with regularization. In contrast to Mészáros, we take the perspective that the regularization of the Newton system leads to an inexact interior-point method. In this section, we propose a specific methodology to update δ_k at each iteration and show that the resulting interior-point method is globally convergent.

Recall that the direction determined by the regularization differs from the true Newton direction in that (15) replaces (4c). Expressed differently, we allow the Newton direction to satisfy

$$E^T \Delta y^k = r_{d_2}^k + \delta_k \Delta z^k,$$

even though we hope that $E^T \Delta y^k$ could equal $r_{d_2}^k$. So the direction given by the regularization is an inexact direction. But what effect does the inexact direction have on convergence? If one tries to extend the convergence proof of Kojima, Megiddo, and Mizuno, all concepts and proofs extend easily except that we no longer have r_α equal to $(1-\alpha)r$, which in turn causes a direct extension of Proposition 3.4 to fail (see section 3.1).

However, it is not so difficult to repair the proof of Proposition 3.4. The key insight is that the degree of inexactness of the direction needs to be controlled in a certain manner. Specifically, if we have

$$\delta_k \|\Delta z^k\| \le \beta\sigma\mu_k/2 \tag{18}$$

for all k, then global convergence is established by Theorem 5.1 below.

Our method for enforcing (18) is essentially to decrease δ_k if (18) does not hold. The resulting algorithm is stated as Algorithm 2. Algorithm 2 incorporates all of the features of Algorithm 1, while adding steps to handle δ_k that are fairly straightforward. It is worth mentioning three items:

- In the `while` loop, each time δ_k is updated, it is decreased by a factor of at least 2. As a result, the `while` loop will terminate after a finite (usually quite small) number of loops.

Algorithm 2. Infeasible Path-Following Algorithm (with Regularization).

Let $\varepsilon > 0$, $\omega > 0$, $\sigma \in (0, 0.99)$, $\delta_1 > 0$, and $(x^1, z^1, y^1, s^1) \in \mathcal{N}$ be given.

for $k = 1, 2, 3, \ldots$ **do**
  If $\mu_k \le \varepsilon$ or $\|(x^k, s^k)\|_1 \ge \omega$, then stop.
  Set ACCEPT $= 0$.
  while ACCEPT $= 0$ **do**
    Set $\nu_k = \sigma\mu_k$ and solve (4) with (4c) replaced by (15) for $(\Delta x^k, \Delta z^k, \Delta y^k, \Delta s^k)$.
    if $\delta_k \|\Delta z^k\| \le \beta\sigma\mu_k/2$ **then**
      ACCEPT $= 1$
    else
      $\delta_k \leftarrow \frac{1}{2} \cdot \beta\sigma\mu_k/(2\|\Delta z^k\|)$
    end if
  end while
  Set $(x^{k+1}, z^{k+1}, y^{k+1}, s^{k+1}) = (x_{\alpha_k}, z_{\alpha_k}, y_{\alpha_k}, s_{\alpha_k})$, where $\alpha_k \in (0, 1]$ is the largest step size such that the relations

$$(x_\alpha, z_\alpha, y_\alpha, s_\alpha) \in \mathcal{N}, \tag{19}$$

$$\mu_\alpha \le (1 - 0.01\alpha)\mu_k \tag{20}$$

  hold for every $\alpha \in [0, \alpha_k]$.
  Set $\delta_{k+1} \leftarrow \beta\sigma\mu_{k+1}/(2\|\Delta z^k\|)$.
end for

- In each loop of the `while` loop, the direction must be recalculated. This necessitates reforming and refactoring the matrix $AXS^{-1}A^T + \delta_k^{-1}EE^T$ of (17) because of the dependence on δ_k. This is a potential downside of Algorithm 2. However, the computational results of section 6 (particularly Table 5) show that, overall, this extra work does not constitute a disadvantage, most likely because the `while` loop is repeated only a small number of times.
- The purpose of the `while` loop is to drive δ_k lower and lower until (18) holds. Based on the (ultimately flawed) idea that δ_k should never increase during the course of the algorithm, our initial implementation of Algorithm 2 maintained a nonincreasing sequence $\{\delta_k\}$. We found by experimentation, however, that sometimes δ_k would go to 0 too quickly, causing numerical difficulties. By this we mean, for example, that one iteration would require $\delta_k \leq 10^{-4}$ to enforce (18), while a later iteration would require only $\delta_k \leq 10^{-2}$. For numerical stability, it makes sense to take δ_k as large as possible, which was not allowed by our initial implementation. Hence, our final implementation (i.e., Algorithm 2) allows δ_k to increase via the last line of the `for` loop, which sets δ_{k+1} to our best guess given current information.

Regarding convergence, the following result holds for this algorithm.

THEOREM 5.1. *Let $\gamma \in (0,1)$, $\beta > 0$, and an initial point $(x^1, z^1, y^1, s^1) \in \mathcal{N}(\gamma, \beta)$ be given. Suppose, moreover, that positive tolerances ε and ω, a dampening factor $\sigma \in (0, 0.99)$, and an initial regularization parameter $\delta_1 > 0$ are specified. Then Algorithm 2 eventually generates an iterate $(x^k, z^k, y^k, s^k) \in \mathcal{N}(\gamma, \beta)$ such that $\mu_k \leq \varepsilon$ or $\|(x^k, s^k)\|_1 \geq \omega$. If the first case occurs, then $\|r^k\| \leq \beta\,\varepsilon$ as well.*

The proof of Theorem 5.1 follows the same steps as that of Theorem 3.1 with two changes: r_α and α_* are now given by

$$r_\alpha = (1-\alpha)r - \alpha\delta\Delta z \tag{21}$$

and

$$\alpha_* := \min\left\{1, \frac{(1-\gamma)\sigma\varepsilon}{\eta}, \frac{\sigma\varepsilon}{2\eta}, \frac{(0.99-\sigma)\varepsilon}{\eta}\right\},$$

respectively, and Proposition 3.4 is replaced by the following proposition.

PROPOSITION 5.2. *If $\delta_k\|\Delta z^k\| \leq \beta\sigma\mu_k/2$, then $\|r_\alpha\| \leq \beta\mu_\alpha$ for all $\alpha \leq \sigma\varepsilon/2\eta$.*

Proof. We have

$$\begin{aligned}
&\beta\mu_\alpha - \|r_\alpha\| \\
&\quad = \beta\left[(1-(1-\sigma)\alpha)\mu + \alpha^2\Delta x^T\Delta s/n\right] - \|(1-\alpha)r - \alpha\delta\Delta z\| \\
&\quad \geq \beta\left[(1-(1-\sigma)\alpha)\mu + \alpha^2\Delta x^T\Delta s/n\right] - (1-\alpha)\|r\| - \alpha\delta\|\Delta z\| \\
&\quad \geq \beta\left[(1-(1-\sigma)\alpha)\mu + \alpha^2\Delta x^T\Delta s/n\right] - (1-\alpha)\|r\| - \beta\alpha\sigma\mu/2 \\
&\quad = (1-\alpha)(\beta\mu - \|r\|) + \beta\alpha\sigma\mu/2 + \beta\alpha^2\Delta x^T\Delta s/n \\
&\quad \geq \beta\alpha\sigma\mu/2 - \beta\alpha^2\eta \geq \alpha\beta\left[\sigma\varepsilon/2 - \alpha\eta\right],
\end{aligned}$$

where the first equality follows from (14) and (21), the first inequality follows from $(x,z,y,s) \in \mathcal{N}$ and (11b), and the second inequality follows from the assumption $(x,z,y,s) \in \mathcal{N}^*$. This proves the result. ∎

6. Implementation and computational results. We compare our proposed regularization (hereafter denoted REGULARIZE) with

- explicitly solving the indefinite system (7) (EXPLICIT),
- splitting the free variables into the difference of two nonnegative variables (SPLIT), and
- putting the free variables into a second-order cone (QCONE).

The comparisons are carried out using SDPT3 (version 3.02), which supports EXPLICIT by default and supports SPLIT and QCONE via modified data input. Moreover, SDPT3 requires only simple code modifications to implement REGULARIZE. All tests are run on a dual Opteron 2.8 GHz, using the HKM direction and the predictor-corrector option.

We point out that, since our implementation is based on SDPT3, REGULARIZE differs from Algorithm 2 in the same ways that typical implementations of interior-point methods differ from theory. For example, the iterates are not explicitly forced to stay in a neighborhood, and Mehrotra's predictor-corrector method is used. The lack of explicit neighborhood does have an impact on how we enforce condition (18). In each iteration, (18) is enforced with the definition $\beta := \mu/\|r\|$. This can be interpreted as assigning to β the smallest value such that the current iterate is actually a member of $\mathcal{N}$ (if membership in $\mathcal{N}$ were maintained). In this sense, the choice of β is also conservative in that it results in the strictest realization of (18). One additional implementation detail: The parameter δ_k is initialized to $\delta_0 = \mu_0$.

As mentioned previously, SDPT3 has been chosen for several reasons, including the fact that its fundamental algorithm closely matches the algorithmic framework of our analysis. As a consequence, we caution that the conclusions supported by our computational results do not necessarily say anything about the effect of regularization within other codes (although we are optimistic that the regularization can have benefits elsewhere; see section 7).

We also note that SDPT3 offers the option of handling dense columns of (A, E) in such a way that computational effort is minimized. We have tested our regularization with this option enabled (which is the default) or disabled, and the method works just as well in both cases.

There do not appear to be many existing test instances of linear conic optimization problems with free variables. For example, the commonly used DIMACS set of benchmark problems [22] contains only 9 instances with free variables, while SDPLIB [3] contains none. We include the 9 DIMACS test problems in our experiments. Kobayashi, Nakata, and Kojima [9] generated modified problems having free variables from SDPLIB to test their approach; starting with the same SDPLIB problems, we generated our own sparse versions of these problems having free variables via a random matrix $E \in \Re^{m\times p}$, where $p := m/2$. In fact, we generated two sets of modified SDPLIB problems: one set with $\text{rank}(E) = p$ and a second set with $\text{rank}(E) = p/2$. In contrast to the theoretical assumption in section 2 that E has full-column rank, we test instances with E having small-column rank because in practice E may have (nearly) dependent columns.

On the other hand, problems with free variables have become very relevant due to recent applications of SDP to certain classes of problems. In particular, we report test results on two additional sets of problems: one from quantum chemistry and another obtained by generating moment relaxations of combinatorial optimization problems [13] using YALMIP [14]. The set of moment relaxations consists of (small) randomly generated maximum-cut, quadratic-knapsack, and stable-set instances. Half of the underlying instances have 15 variables; the other half have 17. The moment

TABLE 1
Overview of the test problem sets.

Class	# of instances	K
DIMACS challenge	9	LP+SOCP
Modified SDPLIB ($\text{rank}(E) = p$)	27	SDP
Modified SDPLIB ($\text{rank}(E) = p/2$)	27	SDP
Quantum chemistry	12	SDP
Moment relaxations	60	SDP

TABLE 2
Characteristics of the test problem sets.

	p			n			m		
Class	min	med	max	min	med	max	min	med	max
DIMACS challenge	1	2	7201	2379	4191	261364	123	3680	130141
Modified SDPLIB ($\text{rank}(E) = p$)	52	125	1514	351	7750	31375	104	250	3028
Modified SDPLIB ($\text{rank}(E) = p/2$)	52	125	1514	351	7750	31375	104	250	3028
Quantum chemistry	35	69	95	38865	480459	1356933	465	2354	4743
Moment relaxations	136	2040	2907	9316	13031	17613	3875	4930	5984

TABLE 3
Accuracy of each method, averaged over instances in each set.

Class	Reg	Exp	Split	qCone
DIMACS challenge	-10.3	-10.4	-5.6	-7.6
Modified SDPLIB ($\text{rank}(E) = p$)	-7.8	-7.9	-4.2	-4.5
Modified SDPLIB ($\text{rank}(E) = p/2$)	-7.9	1.3	-4.4	-4.7
Quantum chemistry	-5.1	-2.1	-1.8	-2.0
Moment relaxations	-6.0	-2.0	-2.5	-3.4

relaxations are of order 2. For the maximum-cut and stable-set instances, we wrote our own simple random generation procedure, whereas the quadratic-knapsack problems were generated as in the study by Caprara, Pisinger, and Toth [4]. The stable-set formulation is due to Motzkin and Straus [20]. The large number of free variables in these instances arises from the number of equality constraints in the original problems. Characteristics of the test problems that we use are summarized in Tables 1 and 2.

The main criterion for comparing the various methods is the resulting accuracy; i.e., we wish to determine which approach yields the most accurate solutions. Specifically, we report accuracies as $\log_{10}$ of the maximum of the standard DIMACS accuracy errors [18]. Roughly speaking, an accuracy of $-k$ in our results corresponds to k digits of accuracy in the reported optimal value.

It is important to point out that SDPT3 tries to the improve the accuracy of the solution from iteration to iteration, until the accuracy deteriorates for a few iterations, at which point SDPT3 stops. For all of the approaches, we let SDPT3 run until this happens and report the best accuracy obtained overall.

The results of our computational experiments are summarized in Tables 3, 4, and 5. Table 3 shows that, for the test sets *DIMACS challenge* and *modified SDPLIB*, the proposed regularization approach basically matches the best accuracy among the other three approaches. More interestingly, the accuracy obtained for the three other test sets is significantly higher. The results for the moment relaxations are particularly interesting because these instances contain the largest proportions of free variables.

TABLE 4
Number of iterations of each method, averaged over instances in each set.

Class	Reg	Exp	Split	qCone
DIMACS challenge	29.7	31.3	17.1	32.0
Modified SDPLIB (rank$(E) = p$)	15.6	15.4	13.0	13.3
Modified SDPLIB (rank$(E) = p/2$)	14.5	11.1	12.1	12.4
Quantum chemistry	22.7	10.0	11.8	14.0
Moment relaxations	21.9	10.8	13.2	16.0

TABLE 5
Average CPU time per iteration (as a percentage of REGULARIZE *time).*

Class	Reg	Exp	Split	qCone
(all)	100%	138%	107%	114%

Looking at the results for the set of problems with E having linearly dependent columns, we note that EXPLICIT achieves very poor accuracy for these problems due to numerical difficulties. The other three approaches, including REGULARIZE, seem unaffected by the dependencies in E.

Table 4 shows that the higher accuracy obtained by REGULARIZE typically requires a higher number of iterations (the other methods stop when accuracy deteriorates). Nonetheless, with respect to CPU time, Table 5 shows that REGULARIZE requires, on average, the same or less computational effort per iteration as that required by the other approaches. In summary, the proposed regularization seems to lead to an overall improvement in the performance of SDPT3 for instances with free variables.

7. Conclusion and future research. We have considered the regularization approach for handling free variables within interior-point methods for conic linear optimization. Using a global convergence analysis, we derive a simple computational strategy for handling and updating the regularization. Straightforward modifications to the modern code SDPT3 allow the regularization to be incorporated within an inexact infeasible primal-dual path-following interior-point framework. Computational results with SDPT3 on a variety of test problems suggest that the regularization is able to achieve the same or significantly better numerical accuracy than other strategies for handling free variables, while requiring less CPU time per iteration on average.

It remains to be studied what impact regularization can have within other SDP codes, including the well-known homogeneous self-dual embedding algorithm implemented in SeDuMi, a package known to yield excellent accuracy in its computations. As indicated in the introduction, our experience with SeDuMi confirms SeDuMi's reputation; SeDuMi achieves 10 to 11 digits of accuracy on the test problems of this paper. It would be interesting to investigate how regularization can be applied to SeDuMi and if it can improve SeDuMi further. Another intriguing possibility is the study of an easily implementable update strategy for the regularization that would permit a proof of polynomial-time convergence.

Finally, we hope that this paper will stimulate further research on techniques to handle conic linear optimization problems with a significant proportion of free variables, as such problems have become relevant in the context of recent applications of SDP to several challenging problems.

Acknowledgment. The authors are in debt to two anonymous referees for numerous insightful comments and suggestions, which have greatly improved the paper.

REFERENCES

[1] E. Andersen, *Handling free variables in primal-dual interior-point methods using a quadratic cone*, in Proceedings of the SIAM Conference on Optimization, Toronto, 2002.

[2] M.F. Anjos, *An improved semidefinite programming relaxation for the satisfiability problem*, Math. Program., 102 (2005), pp. 589–608.

[3] B. Borchers, *SDPLIB* 1.2*, library of semidefinite programming test problems*, Optim. Methods Softw., 11/12 (1999), pp. 683–690.

[4] A. Caprara, D. Pisinger, and P. Toth, *Exact solution of the quadratic knapsack problem*, INFORMS J. Comput., 11 (1999), pp. 125–137.

[5] E. de Klerk, *Aspects of Semidefinite Programming*, Appl. Optim. 65, Kluwer Academic, Dordrecht, 2002.

[6] J. Faraut and A. Korányi, *Analysis on Symmetric Cones*, Oxford Math. Monogr., Oxford University Press, New York, 1994.

[7] R.W. Freund, F. Jarre, and S. Mizuno, *Convergence of a class of inexact interior-point algorithms for linear programs*, Math. Oper. Res., 24 (1999), pp. 50–71.

[8] C. Helmberg, http://www-user.tu-chemnitz.de/~helmberg/semidef.html.

[9] K. Kobayashi, K. Nakata, and M. Kojima, *A Conversion of an SDP Having Free Variables into the Standard Form SDP*, Comput. Optim. Appl., 36 (2007), pp. 289–307.

[10] M. Kojima, N. Megiddo, and S. Mizuno, *A primal-dual infeasible-interior-point algorithm for linear programming*, Math. Program., 61 (1993), pp. 263–280.

[11] J. Korzak, *Convergence analysis of inexact infeasible-interior-point algorithms for solving linear programming problems*, SIAM J. Optim., 11 (2000), pp. 133–148.

[12] J.B. Lasserre, *Global optimization with polynomials and the problem of moments*, SIAM J. Optim., 11 (2001), pp. 796–817.

[13] J.B. Lasserre, *An explicit equivalent positive semidefinite program for nonlinear* 0-1 *programs*, SIAM J. Optim., 12 (2002), pp. 756–769.

[14] J. Löfberg, *YALMIP: A toolbox for modeling and optimization in MATLAB*, in Proceedings of the CACSD Conference, Taipei, Taiwan, 2004. Available from http://control.ee.ethz.ch/~joloef/yalmip.php.

[15] I. Maros and C. Mészáros, *The role of the augmented system in interior point methods*, European J. Oper. Res., 107 (1998), pp. 720–736.

[16] S. Mehrotra, *On the implementation of a primal-dual interior point method*, SIAM J. Optim., 2 (1992), pp. 575–601.

[17] C. Mészáros, *On free variables in interior point methods*, Optim. Methods Softw., 9 (1998), pp. 121–139.

[18] H.D. Mittelmann, *An independent benchmarking of SDP and SOCP solvers*, Math. Program., 95 (2003), pp. 407–430.

[19] S. Mizuno and F. Jarre, *Global and polynomial-time convergence of an infeasible-interior-point algorithm using inexact computation*, Math. Program., 84 (1999), pp. 105–122.

[20] T. S. Motzkin and E. G. Straus, *Maxima for graphs and a new proof of a theorem of Turán*, Canad. J. Math., 17 (1965), pp. 533–540.

[21] P.A. Parrilo, *Semidefinite programming relaxations for semialgebraic problems*, Math. Program., 96 (2003), pp. 293–320.

[22] G. Pataki and S. Schmieta, *The DIMACS Library of Semidefinite-Quadratic-Linear Programs*, Technical report, Rutgers, 1999. See http://dimacs.rutgers.edu/Challenges/Seventh/Instances/.

[23] I. Pólik, *Addendum to the SeDuMi User Guide Version* 1.1, Technical report, Advanced Optimization Laboratory, McMaster University, 2005.

[24] M. Salahi, J. Peng, and T. Terlaky, *On Mehrotra-type Predictor-Corrector Algorithms*, AdvOl-Report No. 2005/4, Advanced Optimization Laboratory, 2005.

[25] J.F. Sturm, *Using SeDuMi* 1.02*, a MATLAB toolbox for optimization over symmetric cones*, Optim. Methods Softw., 11/12 (1999), pp. 625–653.

[26] R. H. Tütüncü, K. C. Toh, and M. J. Todd, *SDPT*3*: A Matlab Software Package for Semidefinite-Quadratic-Linear Programming, Version* 3.0, 2001. Available from http://www.math.nus.edu.sg/~mattohkc/sdpt3.html.

[27] R.J. Vanderbei, *Symmetric quasidefinite matrices*, SIAM J. Optim., 5 (1995), pp. 100–113.

[28] R.J. Vanderbei, *Linear Programming*, Internat. Ser. in Oper. Res. Management Sci. 37, second ed., Kluwer Academic, Boston, MA, 2001.

[29] H. Waki, S. Kim, M. Kojima, and M. Muramatsu, *Sums of Squares and Semidefinite Programming Relaxations for Polynomial Optimization Problems with Structured Sparsity*, SIAM J. Optim., 17 (2006), pp. 218–242.

[30] H. Wolkowicz, R. Saigal, and L. Vandenberghe, eds., *Handbook of Semidefinite Programming*, Kluwer Academic, Boston, MA, 2000.
[31] S.J. Wright, *Primal-Dual Interior-Point Methods*, SIAM, Philadelphia, PA, 1997.
[32] Y. Zhang, *On the convergence of a class of infeasible interior-point methods for the horizontal linear complementarity problem*, SIAM J. Optim., 4 (1994), pp. 208–227.
[33] Y. Zhang, *On extending some primal-dual interior-point algorithms from linear programming to semidefinite programming*, SIAM J. Optim., 8 (1998), pp. 365–386.
[34] Y. Zhang, *User's guide to LIPSOL: linear-programming interior point solvers V*0.4, Optim. Methods Softw., 11/12 (1999), pp. 385–396.
[35] Y. Zhang and De T. Zhang, *On polynomiality of the Mehrotra-type predictor-corrector interior-point algorithms*, Math. Program., 68 (1995), pp. 303–318.
[36] Z. Zhao, B.J. Braams, M. Fukuda, M.L. Overton, and J.K. Percus, *The reduced density matrix method for electronic structure calculations and the role of three-index representability*, J. Chem. Phys., 120 (2004), pp. 2095–2104.
[37] G. Zhou and K.-C. Toh, *Polynomiality of an inexact infeasible interior point algorithm for semidefinite programming*, Math. Program., 99 (2004), pp. 261–282.

SIAM J. OPTIM.
Vol. 18, No. 4, pp. 1326–1350

EXACT REGULARIZATION OF CONVEX PROGRAMS*

MICHAEL P. FRIEDLANDER† AND PAUL TSENG‡

Abstract. The regularization of a convex program is *exact* if all solutions of the regularized problem are also solutions of the original problem for all values of the regularization parameter below some positive threshold. For a general convex program, we show that the regularization is exact if and only if a certain selection problem has a Lagrange multiplier. Moreover, the regularization parameter threshold is inversely related to the Lagrange multiplier. We use this result to generalize an exact regularization result of Ferris and Mangasarian [*Appl. Math. Optim.*, 23 (1991), pp. 266–273] involving a linearized selection problem. We also use it to derive necessary and sufficient conditions for exact penalization, similar to those obtained by Bertsekas [*Math. Programming*, 9 (1975), pp. 87–99] and by Bertsekas, Nedić, and Ozdaglar [*Convex Analysis and Optimization*, Athena Scientific, Belmont, MA, 2003]. When the regularization is not exact, we derive error bounds on the distance from the regularized solution to the original solution set. We also show that existence of a "weak sharp minimum" is in some sense close to being necessary for exact regularization. We illustrate the main result with numerical experiments on the ℓ_1 regularization of benchmark (degenerate) linear programs and semidefinite/second-order cone programs. The experiments demonstrate the usefulness of ℓ_1 regularization in finding sparse solutions.

Key words. convex program, conic program, linear program, regularization, exact penalization, Lagrange multiplier, degeneracy, sparse solutions, interior-point algorithms

AMS subject classifications. 90C25, 90C05, 90C51, 65K10, 49N15

DOI. 10.1137/060675320

1. Introduction.

A common approach to solving an ill-posed problem—one whose solution is not unique or is acutely sensitive to data perturbations—is to construct a related problem whose solution is well behaved and deviates only slightly from a solution of the original problem. This is known as regularization, and deviations from solutions of the original problem are generally accepted as a trade-off for obtaining solutions with other desirable properties. However, it would be more desirable if solutions of the regularized problem were also solutions of the original problem. Here we present necessary and sufficient conditions for this to hold and study their implications for general convex programs.

Consider the general convex program

$$\text{(P)} \qquad \begin{array}{ll} \underset{x}{\text{minimize}} & f(x) \\ \text{subject to} & x \in \mathcal{C}, \end{array}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function, and $\mathcal{C} \subseteq \mathbb{R}^n$ is a nonempty closed convex set. In cases where (P) is ill-posed or lacks a smooth dual, a popular technique is to regularize the problem by adding a convex function to the objective. This yields the

*Received by the editors November 18, 2006; accepted for publication April 17, 2007; published electronically November 14, 2007.

http://www.siam.org/journals/siopt/18-4/67532.html

†Department of Computer Science, University of British Columbia, Vancouver V6T 1Z4, BC, Canada (mpf@cs.ubc.ca). This author's research was supported by the National Science and Engineering Council of Canada.

‡Department of Mathematics, University of Washington, Seattle, WA 98195 (tseng@math.washington.edu). This author's research was supported by National Science Foundation grant DMS-0511283.

regularized problem

$$(\mathrm{P}_\delta) \qquad \begin{array}{ll} \underset{x}{\text{minimize}} & f(x) + \delta\phi(x) \\ \text{subject to} & x \in \mathcal{C}, \end{array}$$

where $\phi : \mathbb{R}^n \to \mathbb{R}$ is a convex function, and δ is a nonnegative regularization parameter. The regularization function ϕ may be nonlinear and/or nondifferentiable.

In general, solutions of the regularized problem (P_δ) need not be solutions of (P). (Here and throughout, "solution" is used in lieu of "optimal solution.") We say that the regularization is *exact* if the solutions of (P_δ) are also solutions of (P) for all values of δ below some positive threshold value $\bar{\delta}$. We choose the term *exact* to draw an analogy with exact penalization that is commonly used for solving constrained nonlinear programs. An exact penalty formulation of a problem can recover a solution of the original problem for all values of the penalty parameter beyond a threshold value. See, for example, [4, 5, 9, 21, 24, 31] and, for more recent discussions, [7, 15].

Exact regularization can be useful for various reasons. If a convex program does not have a unique solution, exact regularization may be used to select solutions with desirable properties. In particular, Tikhonov regularization [45], which corresponds to $\phi(x) = \|x\|_2^2$, can be used to select a least two-norm solution. Specialized algorithms for computing least two-norm solutions of linear programs (LPs) have been proposed by [25, 26, 27, 30, 33, 48], among others. Saunders [42] and Altman and Gondzio [1] use Tikhonov regularization as a tool for influencing the conditioning of the underlying linear systems that arise in the implementation of large-scale interior-point algorithms for LPs. Bertsekas [4, Proposition 4] and Mangasarian [30] use Tikhonov regularization to form a smooth convex approximation of the dual LP.

More recently, there has been much interest in ℓ_1 regularization, which corresponds to $\phi(x) = \|x|_1$. Recent work related to signal processing has focused on using LPs to obtain sparse solutions (i.e., solutions with few nonzero components) of underdetermined systems of linear equations $Ax = b$ (with the possible additional condition $x \geq 0$); for examples, see [13, 12, 14, 18]. In machine learning and statistics, ℓ_1 regularization of linear least-squares problems (sometimes called lasso regression) plays a prominent role as an alternative to Tikhonov regularization; for examples, see [19, 44]. Further extensions to regression and maximum likelihood estimation are studied in [2, 41], among others.

There have been some studies of exact regularization for the case of differentiable ϕ, mainly for LP [4, 30, 34], but to our knowledge there has been only one study, by Ferris and Mangasarian [20], for the case of nondifferentiable ϕ. However, their analysis is mainly for the case of strongly convex ϕ, and thus is not applicable to regularization functions such as the one-norm. In this paper, we study exact regularization of the convex program (P) by (P_δ) for a general convex ϕ.

Central to our analysis is a related convex program that selects solutions of (P) of least ϕ-value:

$$(\mathrm{P}^\phi) \qquad \begin{array}{lll} \underset{x}{\text{minimize}} & \phi(x) & \\ \text{subject to} & x \in \mathcal{C}, & f(x) \leq p^*, \end{array}$$

where p^* denotes the optimal value of (P). We assume a nonempty solution set of (P), which we denote by $\mathcal{S}$, so that p^* is finite and (P^ϕ) is feasible. Clearly, any solution of (P^ϕ) is also a solution of (P). The converse, however, does not generally hold.

In section 2 we prove our main result: the regularization (P_δ) is exact if and only if the selection problem (P^ϕ) has a Lagrange multiplier μ^*. Moreover, the solution set of (P_δ) coincides with the solution set of (P^ϕ) for all $\delta < 1/\mu^*$; see Theorem 2.1 and Corollary 2.2.

A particular case of special interest is conic programs, which correspond to

$$f(x) = c^T x \qquad \text{and} \qquad \mathcal{C} = \{x \in \mathcal{K} \mid Ax = b\}, \tag{1.1}$$

where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$, and $\mathcal{K} \subseteq \mathbb{R}^n$ is a nonempty closed convex cone. In the further case where $\mathcal{K}$ is polyhedral, (P^ϕ) always has a Lagrange multiplier. Thus we extend a result obtained by Mangasarian and Meyer for LPs [34, Theorem 1]; their (weaker) result additionally assumes differentiability (but not convexity) of ϕ on $\mathcal{S}$, and proves that existence of a Lagrange multiplier for (P^ϕ) implies existence of a common solution x^* of (P) and (P_δ) for all positive δ below some threshold. In general, however, (P^ϕ) need not have a Lagrange multiplier even if $\mathcal{C}$ has a nonempty interior. This is because the additional constraint $f(x) = c^T x \le p^*$ may exclude points in the interior of $\mathcal{C}$. We discuss this further in section 2.

1.1. Applications. We present four applications of our main result. The first three show how to extend existing results in convex optimization. The fourth shows how exact regularization can be used in practice.

Linearized selection (section 3). In the case where f is differentiable, $\mathcal{C}$ is polyhedral, and ϕ is strongly convex, Ferris and Mangasarian [20, Theorem 9] show that the regularization (P_δ) is exact if and only if the solution of (P^ϕ) is unchanged when f is replaced by its linearization at any $\bar{x} \in \mathcal{S}$. We generalize this result by relaxing the strong convexity assumption on ϕ; see Theorem 3.2.

Exact penalization (section 4). We show a close connection between exact regularization and exact penalization by applying our main results to obtain necessary and sufficient conditions for exact penalization of convex programs. The resulting conditions are similar to those obtained by Bertsekas [4, Proposition 1], Mangasarian [31, Theorem 2.1], and Bertsekas, Nedić, and Ozdaglar [7, section 7.3]; see Theorem 4.2.

Error bounds (section 5). We show that in the case where f is continuously differentiable, $\mathcal{C}$ is polyhedral, and $\mathcal{S}$ is bounded, a necessary condition for exact regularization with any ϕ is that f has a "weak sharp minimum" [10, 11] over $\mathcal{C}$. In the case where the regularization is not exact, we derive error bounds on the distance from each solution of the regularized problem (P_δ) to $\mathcal{S}$ in terms of δ and the growth rate of f on $\mathcal{C}$ away from $\mathcal{S}$.

Sparse solutions (section 6). As an illustration of our main result, we apply exact ℓ_1 regularization to select sparse solutions of conic programs. In section 6.1 we report numerical results on a set of benchmark LPs from the NETLIB [36] test set and on a set of randomly generated LPs with prescribed dual degeneracy (i.e., nonunique primal solutions). Analogous results are reported in section 6.2 for a set of benchmark semidefinite programs (SDPs) and second-order cone programs (SOCPs) from the DIMACS test set [37]. The numerical results highlight the effectiveness of this approach for inducing sparsity in the solutions obtained via an interior-point algorithm.

1.2. Assumptions. The following assumptions hold implicitly throughout.

Assumption 1.1 (feasibility and finiteness). The feasible set $\mathcal{C}$ is nonempty and the solution set $\mathcal{S}$ of (P) is nonempty.

Assumption 1.2 (bounded level sets). The level set $\{x \in \mathcal{S} \mid \phi(x) \le \beta\}$ is bounded for each $\beta \in \mathbb{R}$, and $\inf_{x\in\mathcal{C}} \phi(x) > -\infty$. (For example, this assumption holds when ϕ is coercive.)

Assumption 1.1 implies that the optimal value p^* of (P) is finite. Assumptions 1.1 and 1.2 together ensure that the solution set of (P^ϕ), denoted by $\mathcal{S}^\phi$, is nonempty and compact, and that the solution set of (P_δ), denoted by $\mathcal{S}_\delta$, is nonempty and compact for all $\delta > 0$. The latter is true because, for any $\delta > 0$ and $\beta \in \mathbb{R}$, any point x in the level set $\{x \in \mathcal{C} \mid f(x) + \delta\phi(x) \leq \beta\}$ satisfies $f(x) \geq p^*$ and $\phi(x) \geq \inf_{x' \in \mathcal{C}} \phi(x')$, so that $\phi(x) \leq (\beta - p^*)/\delta$ and $f(x) \leq \beta - \delta \inf_{x' \in \mathcal{C}} \phi(x')$. Assumptions 1.1 and 1.2 then imply that ϕ, f, and $\mathcal{C}$ have no nonzero recession direction in common, so the above level set must be bounded [40, Theorem 8.7].

Our results can be extended accordingly if the above assumptions are relaxed to the assumption that $\mathcal{S}^\phi \neq \emptyset$ and $\mathcal{S}_\delta \neq \emptyset$ for all $\delta > 0$ below some positive threshold.

2. Main results. Ferris and Mangasarian [20, Theorem 7] prove that if the objective function f is linear, then

$$\bigcap_{0<\delta<\bar{\delta}} \mathcal{S}_\delta \subseteq \mathcal{S}^\phi \tag{2.1}$$

for any $\bar{\delta} > 0$. However, an additional constraint qualification on $\mathcal{C}$ is needed to ensure that the set on the left-hand side of (2.1) is nonempty (see [20, Theorem 8]). The following example shows that the set can be empty:

$$\underset{x}{\text{minimize}} \quad x_3 \quad \text{subject to} \quad x \in \mathcal{K}, \tag{2.2}$$

where $\mathcal{K} = \{(x_1, x_2, x_3) \mid x_1^2 \leq x_2 x_3,\ x_2 \geq 0,\ x_3 \geq 0\}$, i.e., $\mathcal{K}$ defines the cone of 2×2 symmetric positive semidefinite matrices. Clearly $\mathcal{K}$ has a nonempty interior, and the solutions have the form $x_1^* = x_3^* = 0$, $x_2^* \geq 0$, with $p^* = 0$. Suppose that the convex regularization function ϕ is

$$\phi(x) = |x_1 - 1| + |x_2 - 1| + |x_3|. \tag{2.3}$$

(Note that ϕ is coercive, but not strictly convex.) Then (P^ϕ) has the singleton solution set $\mathcal{S}^\phi = \{(0, 1, 0)\}$. However, for any $\delta > 0$, (P_δ) has the unique solution

$$x_1 = \frac{1}{2(1 + \delta^{-1})}, \quad x_2 = 1, \quad x_3 = \frac{1}{4(1 + \delta^{-1})^2},$$

which converges to the solution of (P^ϕ) as $\delta \to 0$, but is never equal to it. Therefore $\mathcal{S}_\delta$ differs from $\mathcal{S}^\phi$ for all $\delta > 0$ sufficiently small.

Note that the left-hand side of (2.1) can be empty even when ϕ is strongly convex and infinitely differentiable. As an example, consider the strongly convex quadratic regularization function

$$\phi(x) = |x_1 - 1|^2 + |x_2 - 1|^2 + |x_3|^2.$$

As with (2.3), it can be shown in this case that $\mathcal{S}_\delta$ differs from $\mathcal{S}^\phi = \{(0, 1, 0)\}$ for all $\delta > 0$ sufficiently small. In particular, $(\delta/2, 1, \delta^2/4)$ is feasible for (P_δ), and its objective function value is strictly less than that of $(0, 1, 0)$. Thus the latter cannot be a solution of (P_δ) for any $\delta > 0$.

In general, one can show that as $\delta \to 0$, each cluster point of solutions of (P_δ) belongs to $\mathcal{S}^\phi$. Moreover, there is no duality gap between (P^ϕ) and its dual because $\mathcal{S}^\phi$ is compact (see [40, Theorem 30.4(i)]). However, the supremum in the dual problem

might not be attained, in which case there would be no Lagrange multiplier for (P^ϕ)—and hence no exact regularization property. Thus, additional constraint qualifications are needed when f is not affine or $\mathcal{C}$ is not polyhedral.

The following theorem and corollary are our main results. They show that the regularization (P_δ) is exact if and only if the selection problem (P^ϕ) has a Lagrange multiplier μ^*. Moreover, $\mathcal{S}_\delta = \mathcal{S}^\phi$ for all $\delta < 1/\mu^*$. Parts of our proof bear similarity to the arguments used by Mangasarian and Meyer [34, Theorem 1], who consider the two cases $\mu^* = 0$ and $\mu^* > 0$ separately in proving the "if" direction. However, instead of working with the KKT conditions for (P) and (P_δ), we work with saddle-point conditions.

THEOREM 2.1.
(a) *For any $\delta > 0$, $\mathcal{S} \cap \mathcal{S}_\delta \subseteq \mathcal{S}^\phi$.*
(b) *If there exists a Lagrange multiplier μ^* for (P^ϕ), then $\mathcal{S} \cap \mathcal{S}_\delta = \mathcal{S}^\phi$ for all $\delta \in (0, 1/\mu^*]$.*
(c) *If there exists $\bar{\delta} > 0$ such that $\mathcal{S} \cap \mathcal{S}_{\bar{\delta}} \neq \emptyset$, then $1/\bar{\delta}$ is a Lagrange multiplier for (P^ϕ), and $\mathcal{S} \cap \mathcal{S}_\delta = \mathcal{S}^\phi$ for all $\delta \in (0, \bar{\delta}]$.*
(d) *If there exists $\bar{\delta} > 0$ such that $\mathcal{S} \cap \mathcal{S}_{\bar{\delta}} \neq \emptyset$, then $\mathcal{S}_\delta \subseteq \mathcal{S}$ for all $\delta \in (0, \bar{\delta})$.*

Proof. Part (a). Consider any $x^* \in \mathcal{S} \cap \mathcal{S}_\delta$. Then, because $x^* \in \mathcal{S}_\delta$,

$$f(x^*) + \delta\phi(x^*) \leq f(x) + \delta\phi(x) \quad \text{for all} \quad x \in \mathcal{C}.$$

Also, $x^* \in \mathcal{S}$, so $f(x) = f(x^*) = p^*$ for all $x \in \mathcal{S}$. This implies that

$$\phi(x^*) \leq \phi(x) \quad \text{for all} \quad x \in \mathcal{S}.$$

Thus $x^* \in \mathcal{S}^\phi$, and it follows that $\mathcal{S} \cap \mathcal{S}_\delta \subseteq \mathcal{S}^\phi$.

Part (b). Assume that there exists a Lagrange multiplier μ^* for (P^ϕ). We consider the two cases $\mu^* = 0$ and $\mu^* > 0$ in turn.

First, suppose that $\mu^* = 0$. Then, for any solution x^* of (P^ϕ),

$$x^* \in \arg\min_{x \in \mathcal{C}} \ \phi(x),$$

or, equivalently,

$$\phi(x^*) \leq \phi(x) \quad \text{for all} \quad x \in \mathcal{C}. \tag{2.4}$$

Also, x^* is feasible for (P^ϕ), so $x^* \in \mathcal{S}$. Thus

$$f(x^*) \leq f(x) \quad \text{for all} \quad x \in \mathcal{C}.$$

Multiplying the inequality in (2.4) by $\delta \geq 0$ and adding it to the above inequality yields

$$f(x^*) + \delta\phi(x^*) \leq f(x) + \delta\phi(x) \quad \text{for all} \quad x \in \mathcal{C}.$$

Thus $x^* \in \mathcal{S}_\delta$ for all $\delta \in [0, \infty)$.

Second, suppose that $\mu^* > 0$. Then, for any solution x^* of (P^ϕ),

$$x^* \in \arg\min_{x \in \mathcal{C}} \ \phi(x) + \mu^*(f(x) - p^*),$$

or, equivalently,

$$x^* \in \arg\min_{x\in\mathcal{C}} \; f(x) + \frac{1}{\mu^*}\phi(x).$$

Thus

$$f(x^*) + \frac{1}{\mu^*}\phi(x^*) \le f(x) + \frac{1}{\mu^*}\phi(x) \quad \text{for all} \quad x \in \mathcal{C}.$$

Also, x^* is feasible for (P^ϕ), so that $x^* \in \mathcal{S}$. Therefore

$$f(x^*) \le f(x) \quad \text{for all} \quad x \in \mathcal{C}.$$

Then, for any $\lambda \in [0,1]$, multiplying the above two inequalities by λ and $1-\lambda$, respectively, and summing them yields

$$f(x^*) + \frac{\lambda}{\mu^*}\phi(x^*) \le f(x) + \frac{\lambda}{\mu^*}\phi(x) \quad \text{for all} \quad x \in \mathcal{C}.$$

Thus $x^* \in \mathcal{S}_\delta$ for all $\delta \in [0, 1/\mu^*]$.

The above arguments show that $\mathcal{S}^\phi \subseteq \mathcal{S}_\delta$ for all $\delta \in [0, 1/\mu^*]$, and therefore $\mathcal{S}^\phi \subseteq \mathcal{S}\cap\mathcal{S}_\delta$ for all $\delta \in (0, 1/\mu^*]$. By Part (a) of the theorem, we must have $\mathcal{S}^\phi = \mathcal{S}\cap\mathcal{S}_\delta$ as desired.

Part (c). Assume that there exists $\bar\delta > 0$ such that $\mathcal{S} \cap \mathcal{S}_{\bar\delta} \neq \emptyset$. Then, for any $x^* \in \mathcal{S} \cap \mathcal{S}_{\bar\delta}$, we have $x^* \in \mathcal{S}_{\bar\delta}$, and thus

$$x^* \in \arg\min_{x\in\mathcal{C}} \; f(x) + \bar\delta\phi(x),$$

or, equivalently,

$$x^* \in \arg\min_{x\in\mathcal{C}} \; \phi(x) + \frac{1}{\bar\delta}(f(x) - p^*).$$

By Part (a), $x^* \in \mathcal{S}^\phi$. This implies that any $x \in \mathcal{S}^\phi$ attains the minimum because $\phi(x) = \phi(x^*)$ and $f(x) = p^*$. Therefore $1/\bar\delta$ is a Lagrange multiplier for (P^ϕ). By Part (b), $\mathcal{S} \cap \mathcal{S}_\delta = \mathcal{S}^\phi$ for all $\delta \in (0, \bar\delta]$.

Part (d). To simplify notation, define $f_\delta(x) = f(x) + \delta\phi(x)$. Assume that there exists a $\bar\delta > 0$ such that $\mathcal{S} \cap \mathcal{S}_{\bar\delta} \neq \emptyset$. Fix any $x^* \in \mathcal{S} \cap \mathcal{S}_{\bar\delta}$. For any $\delta \in (0, \bar\delta)$ and any $x \in \mathcal{C} \setminus \mathcal{S}$, we have

$$f_{\bar\delta}(x^*) \le f_{\bar\delta}(x) \quad \text{and} \quad f(x^*) < f(x).$$

Because $0 < \delta/\bar\delta < 1$, this implies that

$$f_\delta(x^*) = \frac{\delta}{\bar\delta} f_{\bar\delta}(x^*) + \left(1 - \frac{\delta}{\bar\delta}\right) f(x^*) < \frac{\delta}{\bar\delta} f_{\bar\delta}(x) + \left(1 - \frac{\delta}{\bar\delta}\right) f(x) = f_\delta(x).$$

Because $x^* \in \mathcal{C}$, this shows that $x \in \mathcal{C}\setminus\mathcal{S}$ cannot be a solution of (P_δ), and so $\mathcal{S}_\delta \subseteq \mathcal{S}$, as desired. □

Theorem 2.1 shows that existence of a Lagrange multiplier μ^* for (P^ϕ) is necessary and sufficient for exact regularization of (P) by (P_δ) for all $0 < \delta < 1/\mu^*$. Coerciveness

of ϕ on $\mathcal{S}$ is needed only to ensure that $\mathcal{S}^\phi$ is nonempty. If $\delta = 1/\mu^*$, then $\mathcal{S}_\delta$ need not be a subset of $\mathcal{S}$. For example, suppose that

$$n = 1, \quad \mathcal{C} = [0, \infty), \quad f(x) = x, \quad \text{and} \quad \phi(x) = |x - 1|.$$

Then $\mu^* = 1$ is the only Lagrange multiplier for (P^ϕ), but $\mathcal{S}_1 = [0, 1] \not\subseteq \mathcal{S} = \{0\}$. If $\mathcal{S}_\delta$ is a singleton for $\delta \in (0, 1/\mu^*]$, such as when ϕ is strictly convex, then Theorem 2.1(b) and $\mathcal{S}^\phi \neq \emptyset$ imply that $\mathcal{S}_\delta \subseteq \mathcal{S}$.

The following corollary readily follows from Theorem 2.1(b)–(d) and $\mathcal{S}^\phi \neq \emptyset$.

COROLLARY 2.2.
(a) *If there exists a Lagrange multiplier μ^* for (P^ϕ), then $\mathcal{S}_\delta = \mathcal{S}^\phi$ for all $\delta \in (0, 1/\mu^*)$.*
(b) *If there exists $\bar{\delta} > 0$ such that $\mathcal{S}_{\bar{\delta}} = \mathcal{S}^\phi$, then $1/\bar{\delta}$ is a Lagrange multiplier for (P^ϕ), and $\mathcal{S}_\delta = \mathcal{S}^\phi$ for all $\delta \in (0, \bar{\delta}]$.*

2.1. Conic programs. Conic programs (CPs) correspond to (P) with f and $\mathcal{C}$ given by (1.1). They include several important problem classes. LPs correspond to $\mathcal{K} = \mathbb{R}^n_+$ (the nonnegative orthant); SOCPs correspond to

$$\mathcal{K} = \mathcal{K}^{\text{soc}}_{n_1} \times \cdots \times \mathcal{K}^{\text{soc}}_{n_K} \qquad \text{with} \qquad \mathcal{K}^{\text{soc}}_n := \left\{ x \in \mathbb{R}^n \;\middle|\; \sum_{i=1}^{n-1} x_i^2 \le x_n^2, \; x_n \ge 0 \right\}$$

(a product of second-order cones); SDPs correspond to $\mathcal{K} = \mathbb{S}^n_+$ (the cone of symmetric positive semidefinite $n \times n$ real matrices). CPs are discussed in detail in [3, 8, 35, 38], among others.

It is well known that when $\mathcal{K}$ is polyhedral, the selection problem (P^ϕ), with f and $\mathcal{C}$ given by (1.1), must have a Lagrange multiplier [40, Theorem 28.2]. In this important case, Corollary 2.2 immediately yields the following exact-regularization result for polyhedral CPs.

COROLLARY 2.3. *Suppose that f and $\mathcal{C}$ have the form given by (1.1) and that $\mathcal{K}$ is polyhedral. Then there exists a positive $\bar{\delta}$ such that $\mathcal{S}_\delta = \mathcal{S}^\phi$ for all $\delta \in (0, \bar{\delta})$.*

Corollary 2.3 extends [34, Theorem 1], which additionally assumes differentiability (though not convexity) of ϕ on $\mathcal{S}$ and proves a weaker result that there exists a common solution $x^* \in \mathcal{S} \cap \mathcal{S}_\delta$ for all positive δ below some threshold. If $\mathcal{S}$ is furthermore bounded, then an "excision lemma" of Robinson [39, Lemma 3.5] can be applied to show that $\mathcal{S}_\delta \subseteq \mathcal{S}$ for all positive δ below some threshold. This result is still weaker than Corollary 2.3, however.

2.2. Relaxing the assumptions on the regularization function. The assumption that ϕ is coercive on $\mathcal{S}$ and is bounded from below on $\mathcal{C}$ (Assumption 1.2) ensures that the selection problem (P^ϕ) and the regularized problem (P_δ) have solutions. This assumption is preserved under the introduction of slack variables for linear inequality constraints. For example, if $\mathcal{C} = \{x \in \mathcal{K} \mid Ax \le b\}$ for some closed convex set $\mathcal{K}$, $A \in \mathbb{R}^{m \times n}$, and $b \in \mathbb{R}^m$, then

$$\tilde{\phi}(x, s) = \phi(x) \quad \text{with} \quad \tilde{\mathcal{C}} = \{(x, s) \in \mathcal{K} \times [0, \infty)^m \mid Ax + s = b\}$$

also satisfies Assumption 1.2. Here $\tilde{\phi}(x,s)$ depends only on x. Can Assumption 1.2 be relaxed?

Suppose that $\phi(x)$ depends only on a subset of coordinates x_J and is coercive with respect to x_J, where $x_J = (x_j)_{j\in J}$ and $J \subseteq \{1,\dots,n\}$. Using the assumption that (P) has a feasible point x^*, it is readily seen that (P_δ) has a solution with respect to x_J for each $\delta > 0$, i.e., the minimization in (P_δ) is attained at some x_J. For an LP (f linear and $\mathcal{C}$ polyhedral) it can be shown that (P_δ) has a solution with respect to all coordinates of x. However, in general this need not be true, even for an SOCP. An example is

$$n = 3, \quad f(x) = -x_2 + x_3, \quad \mathcal{C} = \left\{ x \mid \sqrt{x_1^2 + x_2^2} \le x_3 \right\}, \quad \text{and} \quad \phi(x) = |x_1 - 1|.$$

Here, $p^* = 0$ (since $\sqrt{x_1^2 + x_2^2} - x_2 \ge 0$ always) and solutions are of the form $(0,\xi,\xi)$ for all $\xi \ge 0$. For any $\delta > 0$, (P_δ) has optimal value of zero (achieved by setting $x_1 = 1$, $x_3 = \sqrt{1 + x_2^2}$, and taking $x_2 \to \infty$) but has no solution. In general, if we define

$$\widehat{f}(x_J) := \min_{(x_j)_{j\notin J} \mid x\in\mathcal{C}} f(x),$$

then it can be shown, using convex analysis results [40], that $\widehat{f}$ is convex and lower semicontinuous—i.e., the epigraph of $\widehat{f}$ is convex and closed. Then (P_δ) is equivalent to

$$\underset{x_J}{\text{minimize}} \ \widehat{f}(x_J) + \delta\phi(x_J),$$

with ϕ viewed as a function of x_J. Thus, we can in some sense reduce this case to the one we currently consider. Note that $\widehat{f}$ may not be real-valued, but this does not pose a problem with the proof of Theorem 2.1.

3. Linearized selection. Ferris and Mangasarian [20] develop a related exact-regularization result for the special case where f is differentiable, $\mathcal{C}$ is polyhedral, and ϕ is strongly convex. They show that (P_δ) is an exact regularization if the solution set of the selection problem (P^ϕ) is unchanged when f is replaced by its linearization at any $\bar{x} \in \mathcal{S}$. In this section we show how Theorem 2.1 and Corollary 2.2 can be applied to generalize this result. We begin with a technical lemma, closely related to some results given by Mangasarian [32].

LEMMA 3.1. *Suppose that f is differentiable on $\mathbb{R}^n$ and constant on the line segment joining two points x^* and $\bar{x}$ in $\mathbb{R}^n$. Then*

$$\nabla f(x^*)^T(x - x^*) = \nabla f(\bar{x})^T(x - \bar{x}) \quad \textit{for all} \quad x \in \mathbb{R}^n. \tag{3.1}$$

Moreover, ∇f is constant on the line segment.

Proof. Because f is convex differentiable and is constant on the line segment joining x^* and $\bar{x}$, $\nabla f(x^*)^T(\bar{x} - x^*) = 0$. Because f is convex,

$$f(y) - f(\bar{x}) = f(y) - f(x^*) \ge \nabla f(x^*)^T(y - x^*) \quad \text{for all} \quad y \in \mathbb{R}^n.$$

Fix any $x \in \mathbb{R}^n$. Taking $y = \bar{x} + \alpha(x - \bar{x})$ with $\alpha > 0$ yields

$$f(\bar{x} + \alpha(x - \bar{x})) - f(\bar{x}) \ge \nabla f(x^*)^T(\bar{x} + \alpha(x - \bar{x}) - x^*) = \alpha\nabla f(x^*)^T(x - \bar{x}).$$

Dividing both sides by α and then taking $\alpha \to 0$ yields in the limit

$$\nabla f(\bar{x})^T(x - \bar{x}) \geq \nabla f(x^*)^T(x - \bar{x}) = \nabla f(x^*)^T(x - x^*). \tag{3.2}$$

Switching $\bar{x}$ and x^* in the above argument yields an inequality in the opposite direction. Thus (3.1) holds, as desired.

By taking $x = \alpha(\nabla f(x^*) - \nabla f(\bar{x}))$ in (3.1) and letting $\alpha \to \infty$, we obtain that $\|\nabla f(x^*) - \nabla f(\bar{x})\|_2^2 = 0$ and hence $\nabla f(x^*) = \nabla f(\bar{x})$. This shows that ∇f is constant on the line segment. □

Suppose that f is differentiable at every $\bar{x} \in \mathcal{S}$, and consider a variant of the selection problem (P^ϕ) in which the constraint is linearized about $\bar{x}$:

$$(\mathrm{P}^{\phi,\bar{x}}) \qquad \begin{array}{ll} \underset{x}{\text{minimize}} & \phi(x) \\ \text{subject to} & x \in \mathcal{C}, \quad \nabla f(\bar{x})^T(x - \bar{x}) \leq 0. \end{array}$$

Lemma 3.1 shows that the feasible set of $(\mathrm{P}^{\phi,\bar{x}})$ is the same for all $\bar{x} \in \mathcal{S}$. Since f is convex, the feasible set of $(\mathrm{P}^{\phi,\bar{x}})$ contains $\mathcal{S}$, which is the feasible set of (P^ϕ). Let $\mathcal{S}^{\phi,\bar{x}}$ denote the solution set of $(\mathrm{P}^{\phi,\bar{x}})$. In general $\mathcal{S}^\phi \neq \mathcal{S}^{\phi,\bar{x}}$. In the case where ϕ is strongly convex and $\mathcal{C}$ is polyhedral, Ferris and Mangasarian [20, Theorem 9] show that exact regularization (i.e., $\mathcal{S}^\phi = \mathcal{S}_\delta$ for all $\delta > 0$ sufficiently small) holds if and only if $\mathcal{S}^\phi = \mathcal{S}^{\phi,\bar{x}}$. By using Theorem 2.1, Corollary 2.2, and Lemma 3.1, we can generalize this result by relaxing the assumption that ϕ is strongly convex.

THEOREM 3.2. *Suppose that f is differentiable on $\mathcal{C}$.*
(a) *If there exists a $\bar{\delta} > 0$ such that $\mathcal{S}_{\bar{\delta}} = \mathcal{S}^\phi$, then*

$$\mathcal{S}^\phi \subseteq \mathcal{S}^{\phi,\bar{x}} \quad \text{for all} \quad \bar{x} \in \mathcal{S}. \tag{3.3}$$

(b) *If $\mathcal{C}$ is polyhedral and (3.3) holds, then there exists a $\bar{\delta} > 0$ such that $\mathcal{S}_\delta = \mathcal{S}^\phi$ for all $\delta \in (0, \bar{\delta})$.*

Proof. Part (a). Suppose that there exists a $\bar{\delta} > 0$ such that $\mathcal{S}_{\bar{\delta}} = \mathcal{S}^\phi$. Then by Corollary 2.2, $\mu^* := 1/\bar{\delta}$ is a Lagrange multiplier for (P^ϕ), and for any $x^* \in \mathcal{S}^\phi$,

$$x^* \in \arg\min_{x \in \mathcal{C}} \ \phi(x) + \mu^* f(x). \tag{3.4}$$

Because ϕ and f are real-valued and convex, x^* and μ^* satisfy the optimality condition

$$0 \in \partial\phi(x^*) + \mu^* \nabla f(x^*) + N_{\mathcal{C}}(x^*).$$

Then x^* satisfies the KKT condition for the linearized selection problem

$$\underset{x \in \mathcal{C}}{\text{minimize}} \quad \phi(x) \quad \text{subject to} \quad \nabla f(x^*)^T(x - x^*) \leq 0, \tag{3.5}$$

and is therefore a solution of this problem. By Lemma 3.1, the feasible set of this problem remains unchanged if we replace $\nabla f(x^*)^T(x - x^*) \leq 0$ with $\nabla f(\bar{x})^T(x - \bar{x}) \leq 0$ for any $\bar{x} \in \mathcal{S}$. Thus $x^* \in \mathcal{S}^{\phi,\bar{x}}$. The choice of x^* was arbitrary, and so $\mathcal{S}^\phi \subseteq \mathcal{S}^{\phi,\bar{x}}$.

Part (b). Suppose that $\mathcal{C}$ is polyhedral and (3.3) holds. By Lemma 3.1, the solution set of $(\mathrm{P}^{\phi,\bar{x}})$ remains unchanged if we replace $\nabla f(\bar{x})^T(x - \bar{x}) \leq 0$ by $\nabla f(x^*)^T(x - x^*) \leq 0$ for any $x^* \in \mathcal{S}^\phi$. The resulting problem (3.5) is linearly constrained

and therefore has a Lagrange multiplier $\bar{\mu} \in \mathbb{R}$. Moreover, $\bar{\mu}$ is independent of x^*. By Corollary 2.2(a), the problem

$$\underset{x \in \mathcal{C}}{\text{minimize}} \ \phi(x) + \mu^* \nabla f(x^*)^T x$$

has the same solution set as (3.5) for all $\mu^* > \bar{\mu}$. The necessary and sufficient optimality condition for this convex program is

$$0 \in \partial\phi(x) + \mu^* \nabla f(x^*) + N_{\mathcal{C}}(x).$$

Because (3.3) holds, x^* satisfies this optimality condition. Thus (3.4) holds for all $\mu^* > \bar{\mu}$ or, equivalently, $x^* \in \mathcal{S}_\delta$ for all $\delta \in (0, 1/\bar{\mu})$. Because $\bar{\mu}$ is independent of x^*, this shows that $\mathcal{S}^\phi \subseteq \mathcal{S}_\delta$ for all $\delta \in (0, 1/\bar{\mu})$. And because $\emptyset \neq \mathcal{S}^\phi \subseteq \mathcal{S}$, it follows that $\mathcal{S} \cap \mathcal{S}_\delta \neq \emptyset$ for all $\delta \in (0, 1/\bar{\mu})$. By Theorem 2.1(a) and (d), $\mathcal{S}_\delta \subseteq \mathcal{S}^\phi$ for all $\delta \in (0, 1/\bar{\mu})$. Therefore $\mathcal{S}_\delta = \mathcal{S}^\phi$ for all $\delta \in (0, 1/\bar{\mu})$. □

In the case where ϕ is strongly convex, $\mathcal{S}^\phi$ and $\mathcal{S}^{\phi,\bar{x}}$ are both singletons, so (3.3) is equivalent to $\mathcal{S}^\phi = \mathcal{S}^{\phi,\bar{x}}$ for all $\bar{x} \in \mathcal{S}$. Thus, when $\mathcal{C}$ is also polyhedral, Theorem 3.2 reduces to [20, Theorem 9]. Note that in Theorem 3.2(b) the polyhedrality of $\mathcal{C}$ is needed only to ensure the existence of a Lagrange multiplier for (3.5) and can be relaxed by assuming an appropriate constraint qualification. In particular, if $\mathcal{C}$ is given by inequality constraints, then it suffices that ($\mathrm{P}^{\phi,\bar{x}}$) has a feasible point that strictly satisfies all nonlinear constraints [40, Theorem 28.2].

Naturally, (3.3) holds if f is linear. Thus Theorem 3.2(b) is false if we drop the polyhedrality assumption on $\mathcal{C}$, as we can find examples of convex coercive ϕ, linear f, and closed convex (but not polyhedral) $\mathcal{C}$ for which exact regularization fails; see example (2.2).

4. Exact penalization. In this section we show a close connection between exact regularization and exact penalization by applying Corollary 2.2 to obtain necessary and sufficient conditions for exact penalization of convex programs. Consider the convex program

$$\underset{x}{\text{minimize}} \quad \phi(x) \quad \text{subject to} \quad x \in \mathcal{C}, \quad g(x) := \big(g_i(x)\big)_{i=1}^m \leq 0, \tag{4.1}$$

where $\phi, g_1, \ldots, g_m$ are real-valued convex functions defined on $\mathbb{R}^n$, and $\mathcal{C} \subseteq \mathbb{R}^n$ is a nonempty closed convex set. The penalized form of (4.1) is

$$\underset{x}{\text{minimize}} \quad \phi(x) + \sigma P(g(x)) \quad \text{subject to} \quad x \in \mathcal{C}, \tag{4.2}$$

where σ is a positive penalty parameter and $P : \mathbb{R}^m \to [0, \infty)$ is a convex function having the property that $P(u) = 0$ if and only if $u \leq 0$; see [7, section 7.3]. A well-known example of such a penalty function is

$$P(u) = \|\max\{0, u\}\|_p, \tag{4.3}$$

where $\|\cdot\|_p$ is the p-norm ($1 \leq p \leq \infty$) [22, section 14.3].

The conjugate and polar functions of P [40, subsections 12, 15] are defined, respectively, by

$$P^*(w) := \sup_u \ w^T u - P(u) \qquad \text{and} \qquad P^\circ(w) := \sup_{u \not\leq 0} \ \frac{w^T u}{P(u)}.$$

Note that $P^\circ(\alpha w) = \alpha P^\circ(w)$ for all $\alpha \geq 0$. For P given by (4.3), $P^\circ(w)$ equals the q-norm of w whenever $w \geq 0$, where $1/p + 1/q = 1$. The following lemma gives key properties of these functions that are implicit in the analysis of [7, section 7.3].

LEMMA 4.1. *Suppose that $P : \mathbb{R}^m \to [0, \infty)$ is a convex function and $P(u) = 0$ if and only if $u \leq 0$. Then*

(a) *$P(u) \leq P(v)$ whenever $u \leq v$; and*

(b) $$P^*(w) \begin{cases} = \infty & \text{if } w \not\geq 0; \\ > 0 & \text{if } w \geq 0 \text{ and } P^\circ(w) > 1; \\ = 0 & \text{if } w \geq 0 \text{ and } P^\circ(w) \leq 1. \end{cases}$$

Proof. Part (a). Fix any $u, v \in \mathbb{R}^m$ with $u < v$, and define

$$\pi(\alpha) := P(u + \alpha(v - u)) \quad \text{for all} \quad \alpha \in \mathbb{R}.$$

We have $u + \alpha(v - u) < 0$ for all $\alpha < 0$ sufficiently negative, in which case $\pi(\alpha) = 0$. Because π is convex, this implies that π is nondecreasing and hence $\pi(0) \leq \pi(1)$—i.e., $P(u) \leq P(v)$. Thus $P(u) \leq P(v)$ whenever $u < v$. Because P is continuous on $\mathbb{R}^m$ [40, Theorem 10.1], this yields $P(u) \leq P(v)$ whenever $u \leq v$.

Part (b). Fix any $w \in \mathbb{R}^m$. If $w_i < 0$ for some $i \in \{1, \ldots, m\}$, then by letting $u_i \to -\infty$ and setting all other components of u to zero, we obtain $w^T u - P(u) = w_i u_i \to \infty$ and thus $P^*(w) = \infty$. If $w \geq 0$ and $P^\circ(w) > 1$, then $w^T u > P(u)$ for some $u \not\leq 0$ and thus $P^*(w) \geq w^T u - P(u) > 0$. If $w \geq 0$ and $P^\circ(w) \leq 1$, then $w^T u \leq 0 = P(u)$ for all $u \leq 0$, and $w^T u \leq P(u)$ for all $u \not\leq 0$, so that $w^T u \leq P(u)$ for all $u \in \mathbb{R}^m$ (with equality holding when $u = 0$). Therefore $P^*(w) = 0$. ∎

THEOREM 4.2. *Suppose that* (4.1) *has a nonempty compact solution set. If there exist Lagrange multipliers y^* for* (4.1)*, then the penalized problem* (4.2) *has the same solution set as* (4.1) *for all $\sigma > P^\circ(y^*)$. Conversely, if* (4.1) *and* (4.2) *have the same solution set for some $\sigma = \mu^* > 0$, then* (4.1) *and* (4.2) *have the same solution set for all $\sigma \geq \mu^*$, and there exists a Lagrange multiplier vector y^* for* (4.1) *with $\mu^* \geq P^\circ(y^*)$.*

Proof. Set $f(x) = P\big(g(x)\big)$ for all $x \in \mathbb{R}^n$. By the convexity of $g_1, \ldots, g_m$, P, and Lemma 4.1(a), f is a convex function and thus (4.2) is a convex program. Moreover, any feasible point x^* of (4.1) is a solution of (P) with optimal value $p^* = 0$. Accordingly, we identify (4.2) with (P_δ) (where ϕ is the regularization function and $\delta = 1/\sigma$ is the regularization parameter), and we identify the problem

$$\underset{x}{\text{minimize}} \quad \phi(x) \quad \text{subject to} \quad x \in \mathcal{C}, \quad P\big(g(x)\big) \leq 0 \tag{4.4}$$

with (P^ϕ). Assumptions 1.1 and 1.2 are satisfied because (4.1) has a nonempty compact solution set.

A primal-dual solution pair (x^*, y^*) of (4.1) satisfies the KKT conditions

$$0 \in \partial\phi(x) + \sum_{i=1}^m y_i \partial g_i(x) + N_{\mathcal{C}}(x), \quad y \geq 0, \quad g(x) \leq 0, \quad y^T g(x) = 0. \tag{4.5}$$

By [40, Theorem 23.5], the subdifferential of P at u has the expression $\partial P(u) = \{w \mid w^T u = P(u) + P^*(w)\}$. If $u \leq 0$, then $P(u) = 0$ and, by Lemma 4.1(b), $w^T u = P^*(w)$

only if $w \geq 0$ and $P^\circ(w) \leq 1$. This implies that

$$\partial P(u) = \{w \mid w \geq 0,\ P^\circ(w) \leq 1,\ w^T u = 0\} \quad \text{for all} \quad u \leq 0.$$

We can then express the KKT conditions for (4.4) as

(4.6)
$$0 \in \partial\phi(x) + \mu \sum_{i=1}^{m} w_i \partial g_i(x) + N_{\mathcal{C}}(x), \quad \left\{ \begin{array}{r} w \geq 0 \\ P^\circ(w) \leq 1 \\ \mu \geq 0 \end{array} \right\}, \quad g(x) \leq 0, \quad w^T g(x) = 0.$$

By scaling μ and w by $P^\circ(w)$ and $1/P^\circ(w)$, respectively, and using the positive homogeneous property of P°, we can without loss of generality assume that $P^\circ(w) = 1$ in (4.6). Then, upon comparing (4.5) and (4.6), we see that they are equivalent in the sense that (x^*, y^*) satisfies (4.5) if and only if (x^*, μ^*) satisfies (4.6), where

$$\mu^* w^* = y^* \quad \text{and} \quad \mu^* = P^\circ(y^*),$$

for some $w^* \geq 0$ with $P^\circ(w^*) = 1$. Note that μ^* is a Lagrange multiplier for (4.4). Therefore, by Corollary 2.2(a), (4.2) and (4.4) have the same solution set for all $\sigma > \mu^* = P^\circ(y^*)$.

Conversely, suppose that (4.2) and (4.4) have the same solution set for $\sigma = \mu^* > 0$. Then (P_δ) and (P^ϕ) have the same solution set for $\delta = 1/\mu^*$. By Corollary 2.2(b), μ^* is a Lagrange multiplier for (P^ϕ), and (P_δ) and (P^ϕ) have the same solution set for all $\delta \in (0, 1/\mu^*]$. Therefore, (4.1) and (4.2) have the same solution set for all $\sigma \geq \mu^*$. Moreover, for any $x^* \in \mathcal{S}^\phi$ there exists a vector w^* such that (x^*, μ^*, w^*) satisfies (4.6), and so $y^* := \mu^* w^*$ is a Lagrange multiplier vector for (4.1) that satisfies $P^\circ(y^*) = \mu^* P^\circ(w^*) \leq \mu^*$. □

We can consider a minimum P°-value Lagrange multiplier vector y^* and, similarly, a minimum exact penalty parameter σ. Theorem 4.2 asserts that these two quantities are equal—that is,

$$\left\{ \begin{array}{ll} \inf & P^\circ(y^*) \\ \text{such that} & y^* \in \mathbb{R}^m \text{ is a Lagrange multiplier for (4.1)} \end{array} \right\} = \left\{ \begin{array}{ll} \inf & \sigma \\ \text{such that} & \text{(4.2) has the same solution set as (4.1)} \end{array} \right\}.$$

Theorem 4.2 shows that the existence of Lagrange multipliers y^* with $P^\circ(y^*) < \infty$ is necessary and sufficient for exact penalization. There has been much study of sufficient conditions for exact penalization; see, e.g., [4], [5, Proposition 4.1], and [9]. The results in [4, Propositions 1 and 2] assume the existence of Lagrange multipliers y^* and, for the case of separable P (i.e., $P(u) = \sum_i P_i(u_i)$), prove necessary and sufficient conditions on P and y^* for exact penalization. For separable P, the condition $P^\circ(y^*) \leq \sigma$ reduces to

$$y_i^* \leq \sigma \lim_{u_i \downarrow 0} \frac{P_i(u_i)}{u_i}, \quad i = 1, \ldots, m, \tag{4.7}$$

as derived in [4, Proposition 1]. A similar result was obtained in [31, Theorem 2.1] for the further special case of $P_i(u_i) = \max\{0, u_i\}$. Thus Theorem 4.2 may be viewed as a generalization of these results. (For the standard quadratic penalty $P_i(u_i) = \max\{0, u_i\}^2$, the right-hand side of (4.7) is zero, so (4.7) holds only if $y_i^* = 0$, i.e., the constraint $g_i(x) \leq 0$ is redundant.)

The results in [9, Corollary 2.5.1 and Theorem 5.3] assume either the linear-independence or Slater constraint qualifications in order to ensure existence of Lagrange multipliers. Theorem 4.2 is partly motivated by and very similar to the necessary and sufficient conditions obtained in [7, Proposition 7.3.1]. The connection with exact regularization, however, appears to be new.

Although our results for exact regularization can be used to deduce results for exact penalization, the reverse direction does not appear possible. In particular, applying exact penalization to the selection problem (P^{ϕ}) yields a penalized problem very different from (P_{δ}).

5. Error bounds and weak sharp minimum. Even when exact regularization cannot be achieved, we can still estimate the distance from $\mathcal{S}_{\delta}$ to $\mathcal{S}$ in terms of δ and the growth rate of f away from $\mathcal{S}$. We study this type of error bound in this section.

THEOREM 5.1.
(a) *For any* $\bar{\delta} > 0$, $\cup_{0<\delta\leq\bar{\delta}}\mathcal{S}_{\delta}$ *is bounded.*
(b) *Suppose that there exist* $\tau > 0, \gamma \geq 1$ *such that*

$$f(x) - p^* \geq \tau \operatorname{dist}(x, \mathcal{S})^{\gamma} \quad \textit{for all} \quad x \in \mathcal{C}, \tag{5.1}$$

where $\operatorname{dist}(x, \mathcal{S}) = \min_{x^* \in \mathcal{S}} \|x - x^*\|_2$. *Then, for any* $\bar{\delta} > 0$ *there exists* $\tau' > 0$ *such that*

$$\operatorname{dist}(x_{\delta}, \mathcal{S})^{\gamma-1} \leq \tau'\delta \quad \textit{for all} \quad x_{\delta} \in \mathcal{S}_{\delta},\ \delta \in (0, \bar{\delta}].$$

Proof. Part (a). Fix any $x^* \in \mathcal{S}$ and any $\bar{\delta} > 0$. For any $\delta \in (0, \bar{\delta}]$ and $x_{\delta} \in \mathcal{S}_{\delta}$,

$$f(x^*) + \delta\phi(x^*) \geq f(x_{\delta}) + \delta\phi(x_{\delta}) \geq f(x^*) + \delta\phi(x_{\delta}),$$

and thus $\phi(x^*) \geq \phi(x_{\delta})$. Using $\phi(x_{\delta}) \geq \inf_{x\in\mathcal{C}} \phi(x)$, we have, similarly, that

$$f(x_{\delta}) \leq f(x^*) + \delta\Big(\phi(x^*) - \inf_{x\in\mathcal{C}} \phi(x)\Big) \leq f(x^*) + \bar{\delta}\Big(\phi(x^*) - \inf_{x\in\mathcal{C}} \phi(x)\Big).$$

This shows that $\cup_{0<\delta\leq\bar{\delta}}\mathcal{S}_{\delta} \subseteq \{x \in \mathcal{C} \mid \phi(x) \leq \beta,\ f(x) \leq \beta\}$ for some $\beta \in \mathbb{R}$. Since ϕ, f, and $\mathcal{C}$ have no nonzero recession direction in common (see Assumptions 1.1 and 1.2), the second set is bounded and therefore so is the first set.

Part (b). For any $\delta > 0$ and $x_{\delta} \in \mathcal{S}_{\delta}$, let $x^*_{\delta} \in \mathcal{S}$ satisfy $\|x_{\delta} - x^*_{\delta}\|_2 = \operatorname{dist}(x_{\delta}, \mathcal{S})$. Then

$$\begin{aligned} f(x^*_{\delta}) + \delta\phi(x^*_{\delta}) &\geq f(x_{\delta}) + \delta\phi(x_{\delta}) \\ &\geq f(x^*_{\delta}) + \tau\|x_{\delta} - x^*_{\delta}\|_2^{\gamma} + \delta\phi(x_{\delta}), \end{aligned}$$

which implies that

$$\tau\|x_{\delta} - x^*_{\delta}\|_2^{\gamma} \leq \delta\big(\phi(x^*_{\delta}) - \phi(x_{\delta})\big).$$

Because ϕ is convex and real-valued,

$$\phi(x_{\delta}) \geq \phi(x^*_{\delta}) + \eta_{\delta}^T(x_{\delta} - x^*_{\delta}) \geq \phi(x^*_{\delta}) - \|\eta_{\delta}\|_2\|x_{\delta} - x^*_{\delta}\|_2,$$

for some $\eta_{\delta} \in \partial\phi(x^*_{\delta})$. Combining the above two inequalities yields

$$\tau\|x_{\delta} - x^*_{\delta}\|_2^{\gamma-1} \leq \delta\|\eta_{\delta}\|_2.$$

By Part (a), x_δ lies in a bounded set for all $\delta > 0$, so x_δ^* lies in a bounded subset of $\mathcal{S}$ for all $\delta > 0$. Then η_δ lies in a bounded set [40, Theorem 24.7], so that $\|\eta_\delta\|_2$ is uniformly bounded. This proves the desired bound. □

Error bounds of the form (5.1) have been much studied, especially in the cases of linear growth ($\gamma = 1$) and quadratic growth ($\gamma = 2$); see [6, 10, 11, 28, 29, 46] and references therein. In general, it is known that (5.1) holds for some $\tau > 0$ and $\gamma \geq 1$ whenever f is analytic and $\mathcal{C}$ is bounded [28, Theorem 2.1].

Theorem 5.1 does not make much use of the convexity of f and ϕ, and it readily extends to nonconvex f and ϕ. In the case of $\gamma = 1$ in (5.1) (i.e., f has a "weak sharp minimum" over $\mathcal{C}$), Theorem 5.1(b) implies that $\operatorname{dist}(x_\delta, \mathcal{S}) = 0$ for all $x_\delta \in \mathcal{S}_\delta$—i.e., $\mathcal{S}_\delta \subseteq \mathcal{S}$, whenever $\delta < 1/\tau'$. In this case, then, $\mathcal{S}_\delta = \mathcal{S}^\phi$ whenever $\delta < 1/\tau'$ and $\mathcal{S}_\delta \neq \emptyset$. This gives another exact-regularization result.

The following result shows that it is nearly necessary for f to have a weak sharp minimum over $\mathcal{C}$ in order for there to be exact regularization by any strongly convex quadratic regularization function.

THEOREM 5.2. *Suppose that f is continuously differentiable on $\mathbb{R}^n$ and $\mathcal{S}$ is bounded. If there does not exist $\tau > 0$ such that* (5.1) *holds with $\gamma = 1$, then either*

(i) *there exists a strongly convex quadratic function of the form $\phi(x) = \|x - \widehat{x}\|_2^2$ ($\widehat{x} \in \mathbb{R}^n$) and a scalar $\bar{\delta} > 0$ for which $\mathcal{S}_\delta \neq \mathcal{S}^\phi$ for all $\delta \in (0, \bar{\delta}]$;*

or

(ii) *for every sequence $x^k \in \mathcal{C} \setminus \mathcal{S}$, $k = 1, 2, \ldots$, satisfying*

$$\frac{f(x^k) - p^*}{\operatorname{dist}(x^k, \mathcal{S})} \to 0, \tag{5.2}$$

and every cluster point (x^*, v^*) of $\{(s^k, \frac{x^k - s^k}{\|x^k - s^k\|_2})\}$, we have $x^* + \alpha v^* \notin \mathcal{C}$ for all $\alpha > 0$, where $s^k \in \mathcal{S}$ satisfies $\|x^k - s^k\|_2 = \operatorname{dist}(x^k, \mathcal{S})$.

If case (ii) *occurs, then $\mathcal{C}$ is not polyhedral, and for any $\bar{x} \in \mathcal{S}$,*

$$\mathcal{S} = \operatorname*{arg\,min}_{x \in \mathcal{C}} \ \nabla f(\bar{x})^T x. \tag{5.3}$$

Proof. Suppose that there does not exist $\tau > 0$ such that (5.1) holds with $\gamma = 1$. Then there exists a sequence $x^k \in \mathcal{C} \setminus \mathcal{S}$, $k = 1, 2, \ldots$, that satisfies (5.2). Let $s^k \in \mathcal{S}$ satisfy $\|x^k - s^k\|_2 = \operatorname{dist}(x^k, \mathcal{S})$. Let $v^k = (x^k - s^k)/\|x^k - s^k\|_2$, so that $\|v^k\|_2 = 1$. Because $\mathcal{S}$ is bounded, $\{s^k\}$ is bounded. By passing to a subsequence if necessary, we can assume that (s^k, v^k) converges to some (x^*, v^*). Because s^k is the nearest point projection of x^k onto $\mathcal{S}$, we have $v^k \in N_{\mathcal{S}}(s^k)$, i.e., $(x - s^k)^T v^k \leq 0$ for all $x \in \mathcal{S}$. Taking the limit yields $v^* \in N_{\mathcal{S}}(x^*)$, i.e., $(x - x^*)^T v^* \leq 0$ for all $x \in \mathcal{S}$.

Note that $\{x^k\}$ need not converge to x^* or even be bounded. Now, consider the auxiliary sequence

$$y^k = s^k + \epsilon^k (x^k - s^k) \quad \text{with} \quad \epsilon^k = \frac{1}{\max\{k, \|x^k - s^k\|_2\}},$$

$k = 1, 2, \ldots$. Then $\epsilon^k \in (0, 1]$, $y^k \in \mathcal{C} \setminus \mathcal{S}$, $(y^k - s^k)/\|y^k - s^k\|_2 = v^k$ for all k, and $y^k - s^k \to 0$ (so $y^k \to x^*$). Also, the convexity of f implies $f(y^k) \leq (1 - \epsilon^k) f(s^k) +$

$\epsilon^k f(x^k)$ which, together with $\|y^k - s^k\|_2 = \epsilon^k \|x^k - s^k\|_2$ and $f(s^k) = p^*$, implies

$$0 \le \frac{f(y^k) - f(s^k)}{\|y^k - s^k\|_2} \le \frac{\epsilon^k f(x^k) - \epsilon^k f(s^k)}{\|y^k - s^k\|_2} = \frac{f(x^k) - p^*}{\operatorname{dist}(x^k, \mathcal{S})} \to 0. \tag{5.4}$$

Because $f(y^k) - f(s^k) = \nabla f(s^k)^T (y^k - s^k) + o(\|y^k - s^k\|_2)$ and f is continuously differentiable, (5.4) and $y^k - x^k \to 0$ yield, in the limit,

$$\nabla f(x^*)^T v^* = 0. \tag{5.5}$$

Let $f_\delta(x) = f(x) + \delta\phi(x)$, with

$$\phi(x) = \|x - (x^* + v^*)\|_2^2.$$

Because $v^* \in N_{\mathcal{S}}(x^*)$, we have $\mathcal{S}^\phi = \{x^*\}$.

Suppose that there exists $\alpha > 0$ such that $x^* + \alpha v^* \in \mathcal{C}$. Then, for any $\beta \in (0, \alpha]$,

$$\begin{aligned} f_\delta(x^* + \beta v^*) &= f(x^* + \beta v^*) + \delta\|\beta v^* - v^*\|_2^2 \\ &= f(x^*) + \beta \nabla f(x^*)^T v^* + o(\beta) + \delta(\beta - 1)^2 \|v^*\|_2^2 \\ &= f(x^*) + o(\beta) + \delta(1 - 2\beta + \beta^2) \\ &= f_\delta(x^*) + o(\beta) - \delta\beta(2 - \beta), \end{aligned}$$

where the third equality uses (5.5) and $\|v^*\|_2 = 1$. Thus $x^* + \beta v^* \in \mathcal{C}$ and $f_\delta(x^* + \beta v^*) < f_\delta(x^*)$ for all $\beta > 0$ sufficiently small, implying $\mathcal{S}_\delta \ne \mathcal{S}^\phi$. Therefore, if case (ii) does not occur, then case (i) must occur.

Suppose that case (ii) occurs. First, we claim that, for any $\bar{x} \in \mathcal{S}$,

$$\nabla f(\bar{x})^T (x - \bar{x}) > 0 \quad \text{for all} \quad x \in \mathcal{C} \setminus \mathcal{S}.^1$$

Fix any $\bar{x} \in \mathcal{S}$. Because $\nabla f(\bar{x})^T (x - \bar{x}) = 0$ for all $x \in \mathcal{S}$, this yields (5.3). Next, we claim that $\mathcal{C}$ cannot be polyhedral. If $\mathcal{C}$ were polyhedral, then the minimization in (5.3) would be an LP, for which weak sharp minimum holds. Then there would exist $\tau > 0$ such that

$$\nabla f(\bar{x})^T (x - \bar{x}) \ge \tau \operatorname{dist}(x, \mathcal{S}) \quad \text{for all} \quad x \in \mathcal{C}.$$

Because f is convex and thus $f(x) - p^* = f(x) - f(\bar{x}) \ge \nabla f(\bar{x})^T (x - \bar{x})$ for all $x \in \mathcal{C}$, this would imply that (5.1) holds with $\gamma = 1$, contradicting our assumption. ∎

An example of case (ii) occurring in Theorem 5.2 is

$$n = 2, \quad f(x) = x_2, \quad \text{and} \quad \mathcal{C} = \{x \in \mathbb{R}^2 \mid x_1^2 \le x_2\}.$$

Here $\mathcal{S} = \{(0, 0)\}$, $p^* = 0$, and

$$\frac{f(x) - p^*}{\operatorname{dist}(x, \mathcal{S})} = \frac{x_2}{\|x\|_2} = \frac{1}{\sqrt{(x_1/x_2)^2 + 1}} \quad \text{for all} \quad x \in \mathcal{C} \setminus \mathcal{S}.$$

[1] If this were false, then there would exist $\bar{x} \in \mathcal{S}$ and $x \in \mathcal{C} \setminus \mathcal{S}$ such that $\nabla f(\bar{x})^T (x - \bar{x}) = 0$. (Note that $\nabla f(\bar{x})^T (x - \bar{x}) < 0$ cannot occur because $\bar{x} \in \mathcal{S}$.) Let $s \in \mathcal{S}$ satisfy $\|x - s\|_2 = \operatorname{dist}(x, \mathcal{S})$. By Lemma 3.1, $\nabla f(s)^T (x - s) = 0$. Then for $x^k = s + (x - s)/k$, we would have $x^k \in \mathcal{C} \setminus \mathcal{S}$, $f(x^k) - f(s) = o(1/k)$, and $\operatorname{dist}(x^k, \mathcal{S}) = \|x - s\|_2/k$, so x^k satisfies (5.2) and $s^k = s$ for $k = 1, 2, \ldots$. Because $(s^k, \frac{x^k - s^k}{\|x^k - s^k\|_2}) \to (s, \frac{x - s}{\|x - s\|_2})$ and $s + \alpha \frac{x - s}{\|x - s\|_2} \in \mathcal{C}$ for all $\alpha \in (0, \|x - s\|_2]$, this would contradict case (ii) occurring.

The right-hand side goes to 0 if and only if $x_1/x_2 \to \infty$, in which case $x/\|x\|_2 \to (\pm 1, 0)$, and $\alpha(\pm 1, 0) \notin \mathcal{C}$ for all $\alpha > 0$. Interestingly, we can still find $\phi(x) = \|x - \widehat{x}\|_2^2$ for which $\mathcal{S}_\delta \neq \mathcal{S}^\phi$ for all $\delta > 0$ sufficiently small. For example, take $\phi(x) = (x_1 - 1)^2 + (x_2 - 1)^2$. Then (P_δ) becomes

$$\underset{x}{\text{minimize}} \quad x_2 + \delta(x_1 - 1)^2 + \delta(x_2 - 1)^2 \quad \text{subject to} \quad x_1^2 \leq x_2.$$

It is straightforward to check that $(0, 0)$ does not satisfy the necessary optimality conditions for (P_δ) for all $\delta > 0$. This raises the question of whether case (ii) is subsumed by case (i) when $\mathcal{C}$ is not polyhedral. In section 8, we give an example showing that the answer is "no."

6. Sparse solutions. In this section we illustrate a practical application of Corollary 2.2. Our aim is to find sparse solutions of LPs and CPs that may not have unique solutions. To this end, we let $\phi(x) = \|x\|_1$, which clearly satisfies the required Assumption 1.2. (In general, however, some components of x may be more significant or be at different scales, in which case we may not wish to regularize all components or regularize them equally.)

Regularization based on the one-norm has been used in many applications, with the goal of obtaining sparse or even *sparsest* solutions of underdetermined systems of linear equations and least-squares problems. Some recent examples include [14, 16, 17, 18].

The AMPL model and data files and the MATLAB scripts used to generate all of the numerical results presented in the following subsections can be obtained at http://www.cs.ubc.ca/∼mpf/exactreg/.

6.1. Sparse solutions of linear programs. For underdetermined systems of linear equations $Ax = b$ that arise in fields such as signal processing, the studies in [13, 14, 18] advocate solving

$$\underset{x}{\text{minimize}} \quad \|x\|_1 \quad \text{subject to} \quad Ax = b \quad (\text{and possibly } x \geq 0), \tag{6.1}$$

in order to obtain a sparse solution. This problem can be recast as an LP and be solved efficiently. The sparsest solution is given by minimizing the so-called zero-norm, $\|x\|_0$, which counts the number of nonzero components in x. However, the combinatorial nature of this minimization makes it computationally intractable for all but the simplest instances. Interestingly, there exist reasonable conditions under which a solution of (6.1) is a sparsest solution; see [13, 18].

Following this approach, we use Corollary 2.2 as a guide for obtaining least one-norm solutions of a generic LP,

$$\underset{x}{\text{minimize}} \quad c^T x \quad \text{subject to} \quad Ax = b, \quad l \leq x \leq u, \tag{6.2}$$

by solving its regularized version,

$$\underset{x}{\text{minimize}} \quad c^T x + \delta\|x\|_1 \quad \text{subject to} \quad Ax = b, \quad l \leq x \leq u. \tag{6.3}$$

The vectors l and u are lower and upper bounds on x. In many of the numerical tests given below, the exact ℓ_1 regularized solution of (6.2) (given by (6.3) for small-enough values of δ) is considerably sparser than the solution obtained by solving (6.2) directly. In each instance, we solve the regularized and unregularized problems with

the same interior-point solver. We emphasize that, with an appropriate choice of the regularization parameter δ, the solution of the regularized LP is *also* a solution of the original LP.

We use two sets of test instances in our numerical experiments. The instances of the first set are randomly generated using a degenerate LP generator described in [23]. Those of the second set are derived from the infeasible LPs in the NETLIB collection (http://www.netlib.org/lp/infeas/). Both sets of test instances are further described in subsections 6.1.1–6.1.2.

We follow the same procedure for each test instance. First, we solve the LP (6.2) to obtain an unregularized solution x^* and the optimal value $p^* := c^T x^*$. Next, we solve (P^ϕ), reformulated as an LP, to obtain a Lagrange multiplier μ^* and the threshold value $\bar{\delta} = 1/\mu^*$. Finally, we solve (6.3) with $\delta := \bar{\delta}/2$, reformulated as an LP, to obtain a regularized solution x^*_δ.

We use the log-barrier interior-point algorithm implemented in CPLEX 9.1 to solve each LP. The default CPLEX options are used, except for `crossover = 0` and `comptol = 1e-10`. Setting `crossover = 0` forces CPLEX to use the interior-point algorithm only and to not "cross over" to find a vertex solution. In general, we expect the interior-point algorithm to find the analytic center of the solution set (see [47, Theorems 2.16 and 2.17]), which tends to be less sparse than vertex solutions. The `comptol` option tightens CPLEX's convergence tolerance from its default of `1e-8` to its smallest allowable setting. We do not advocate such a tight tolerance in practice, but the higher accuracy aids in computing the sparsity of a computed solution, which we determine as

$$\|x\|_0 = \mathrm{card}\{x_i \mid |x_i| > \epsilon\}, \tag{6.4}$$

where $\epsilon = 10^{-8}$ is larger than the specified convergence tolerance.

6.1.1. Randomly generated LPs. Six dual-degenerate LPs were constructed using Gonzaga's MATLAB generator [23]. This MATLAB program accepts as inputs the problem size and the dimensions of the optimal primal and dual faces, D_p and D_d, respectively. Gonzaga shows that these quantities must satisfy

$$0 \le D_p \le n - m - 1 \qquad \text{and} \qquad 0 \le D_d \le m - 1. \tag{6.5}$$

The six LPs are constructed with parameters $n = 1000$, $m = 100$, $D_d = 0$, and various levels of D_p set as 0%, 20%, 40%, 60%, 80%, and 100% of the maximum of 899 (given by (6.5)). The instances are, respectively, labeled `random-0`, `random-20`, `random-40`, and so on.

Table 6.1 summarizes the results. We confirm that in each instance the optimal values of the unregularized and regularized problems are nearly identical (at least to within the specified tolerance), so each regularized solution is exact. Except for the "control" instance `random-0`, the regularized solution x^*_δ has a strictly lower one-norm and is considerably sparser than the unregularized solution x^*.

6.1.2. Infeasible LPs. The second set of test instances is derived from a subset of the infeasible NETLIB LPs. For each infeasible LP, we discard the original objective and instead form the problem

$$(\mathrm{P}^{\mathrm{inf}}) \qquad \underset{x}{\text{minimize}} \quad \|Ax - b\|_1 \quad \text{subject to} \quad l \le x \le u,$$

and its regularized counterpart

$$(\mathrm{P}^{\mathrm{inf}}_\delta) \qquad \underset{x}{\text{minimize}} \quad \|Ax - b\|_1 + \delta\|x\|_1 \quad \text{subject to} \quad l \le x \le u.$$

TABLE 6.1

Randomly generated LPs with increasing dimension of the optimal primal face. The arrows indicate differences between values in neighboring columns: → indicates that the value to the right is the same; ↘ indicates that the value to the right is lower; ↗ indicates that the value to the right is larger.

LP	$c^T x^*$	$c^T x^*_\delta$	$\|x^*\|_1$	$\|x^*_\delta\|_1$	$\|x^*\|_0$	$\|x^*_\delta\|_0$	$\bar{\delta}$
random-0	2.5e−13	1.0e−13	9.1e+01	→ 9.1e+01	100	→ 100	1.5e−04
random-20	5.6e−13	6.6e−13	2.9e+02	↘ 2.0e+02	278	↘ 100	2.2e−02
random-40	3.8e−12	3.7e−12	4.9e+02	↘ 2.9e+02	459	↘ 100	2.9e−02
random-60	3.9e−14	9.2e−11	6.7e+02	↘ 3.6e+02	637	↘ 101	3.3e−02
random-80	9.1e−12	8.4e−13	8.9e+02	↘ 4.6e+02	816	↘ 100	2.1e−01
random-100	1.8e−16	3.2e−12	1.0e+03	↘ 5.4e+02	997	↘ 102	1.1e−01

The unregularized problem (P^{inf}) models the plausible situation where we wish to fit a set of infeasible equations in the least one-norm sense. But because the one-norm is not strictly convex or the equations are underdetermined, a solution of (P^{inf}) may not be unique, and the regularized problem (P^{inf}_δ) is used to further select a sparse solution.

The following infeasible NETLIB LPs were omitted because CPLEX returned an error message during the solution of (P^{inf}) or (P^{inf}_δ): `lpi-bgindy`, `lpi-cplex2`, `lpi-gran`, `lpi-klein1`, `lpi-klein2`, `lpi-klein3`, `lpi-qual`, `lpi-refinery`, and `lpi-vol1`.

Table 6.2 summarizes the results. We can see that the regularized solution x^*_δ is exact (i.e., $c^T x^*_\delta = c^T x^*$) and has a one-norm lower than or equal to that of the unregularized solution x^* in all instances. In twelve of the twenty instances, x^*_δ is sparser than x^*. In five of the instances, they have the same sparsity. In three of the instances (`lpi-galenet`, `lpi-itest6`, and `lpi-woodinfe`), x^*_δ is actually less sparse, even though its one-norm is lower.

6.2. Sparse solutions of semidefinite/second-order cone programs. In section 6.1 we used Corollary 2.2 to find sparse solutions of LPs. In this section, we report our numerical experience in finding sparse solutions of SDPs and SOCPs that may not have unique solutions. These are conic programs (P) with f and $\mathcal{C}$ given by (1.1), and $\mathcal{K}$ being the Cartesian product of real space, orthant, second-order cones, and semidefinite cones.

The regularized problem (P_δ) can be put in the conic form

$$\begin{array}{ll} \underset{x,u,v}{\text{minimize}} & c^T x + \delta e^T (u+v) \\ \text{subject to} & Ax = b, \quad x - u + v = 0, \\ & (x,u,v) \in \mathcal{K} \times [0,\infty)^{2n}, \end{array} \tag{6.6}$$

where e is the vector of ones. The selection problem (P^ϕ) can also be put in conic form:

$$\begin{array}{ll} \underset{x,u,v,s}{\text{minimize}} & e^T (u+v) \\ \text{subject to} & Ax = b, \quad x - u + v = 0, \quad c^T x + s = p^*, \\ & (x,u,v,s) \in \mathcal{K} \times [0,\infty)^{2n+1}. \end{array} \tag{6.7}$$

As in section 6.1, we first solve (P) to obtain x^* and the optimal value $p^* := c^T x^*$. Then (6.7) is solved to obtain Lagrange multiplier μ^* and the corresponding threshold value $\bar{\delta} := 1/\mu^*$. Finally, we solve (6.6) with $\delta = \bar{\delta}/2$ to obtain x^*_δ. All three

TABLE 6.2
Least one-norm residual solutions of the infeasible NETLIB LPs.

LP	$c^T x^*$	$c^T x_\delta^*$	$\|x^*\|_1$	$\|x_\delta^*\|_1$	$\|x^*\|_0$	$\|x_\delta^*\|_0$	$\bar{\delta}$
lpi-bgdbg1	3.6e+02	3.6e+02	1.6e+04	↘ 1.3e+04	518	↘ 437	3.3e−03
lpi-bgetam	5.4e+01	5.4e+01	6.0e+03	↘ 5.3e+03	633	↘ 441	3.4e−04
lpi-bgprtr	1.9e+01	1.9e+01	4.7e+03	↘ 3.0e+03	25	↘ 20	3.7e−01
lpi-box1	1.0e+00	1.0e+00	5.2e+02	↘ 2.6e+02	261	→ 261	9.9e−01
lpi-ceria3d	2.5e−01	2.5e−01	8.8e+02	→ 8.8e+02	1780	↘ 1767	6.7e−04
lpi-chemcom	9.8e+03	9.8e+03	1.5e+05	↘ 3.8e+04	711	↘ 591	3.1e−01
lpi-cplex1	3.2e+06	3.2e+06	2.4e+09	↘ 1.5e+09	3811	↘ 3489	1.0e−02
lpi-ex72a	1.0e+00	1.0e+00	4.8e+02	↘ 3.0e+02	215	→ 215	1.6e−01
lpi-ex73a	1.0e+00	1.0e+00	4.6e+02	↘ 3.0e+02	211	→ 211	1.6e−01
lpi-forest6	8.0e+02	8.0e+02	4.0e+05	→ 4.0e+05	54	→ 54	1.2e−03
lpi-galenet	2.8e+01	2.8e+01	1.0e+02	↘ 9.2e+01	10	↗ 11	6.3e−01
lpi-gosh	4.0e−02	4.0e−02	1.5e+04	↘ 7.1e+03	9580	↘ 1075	3.9e−05
lpi-greenbea	5.2e+02	5.2e+02	1.4e+06	↘ 5.6e+05	3658	↘ 1609	1.1e−04
lpi-itest2	4.5e+00	4.5e+00	2.3e+01	→ 2.3e+01	7	→ 7	6.5e−01
lpi-itest6	2.0e+05	2.0e+05	4.8e+05	↘ 4.6e+05	12	↗ 14	4.8e−01
lpi-mondou2	1.7e+04	1.7e+04	3.2e+06	↘ 2.7e+06	297	↘ 244	9.5e−02
lpi-pang	2.4e−01	2.4e−01	1.4e+06	↘ 8.2e+04	536	↘ 336	1.4e−06
lpi-pilot4i	3.3e+01	3.3e+01	6.9e+05	↘ 5.1e+04	773	↘ 627	3.6e−06
lpi-reactor	2.0e+00	2.0e+00	1.5e+06	↘ 1.1e+06	569	↘ 357	4.1e−05
lpi-woodinfe	1.5e+01	1.5e+01	3.6e+03	↘ 2.0e+03	60	↗ 87	5.0e−01

problems—(P), (6.6), and (6.7)—are solved using the MATLAB toolbox SeDuMi (version 1.05) [43], which is a C implementation of a log-barrier primal-dual interior-point algorithm for solving SDP/SOCP. The test instances are drawn from the DIMACS Implementation Challenge library [37], a collection of nontrivial medium-to-large SDP/SOCP arising from applications. We omit those instances for which either (P) is infeasible (e.g., `filtinf1`) or if one of (P), (6.6), or (6.7) cannot be solved because of insufficient memory (e.g., `torusg3-8`). All runs were performed on a PowerPC G5 with 2GB of memory running MATLAB 7.3b.

Table 6.3 summarizes the results. For most of the instances, SeDuMi finds only an inaccurate solution (`info.numerr=1`) for at least one of (P), (6.6), or (6.7). For most instances, however, SeDuMi also finds a value of μ^* that seems reasonable. In some instances (`nb_L2_bessel`, `nql30`, `nql80`, `qssp30`, `qssp60`, `qssp180`, `sch_100_100_scal`, `sch_200_100_scal`, `truss8`), the computed multiplier μ^* is quite large relative to the solution accuracy, and yet $c^T x_\delta^*$ matches $c^T x^*$ in the first three significant digits; this suggests that the regularization is effectively exact. For `nb_L2`, `sch_100_50_scal`, and `sch_100_100_orig`, the discrepancies between $c^T x_\delta^*$ and $c^T x^*$ may be attributed to a SeDuMi numerical failure or primal infeasibility in solving either (P) or (6.7) (thus yielding inaccurate μ^*) or (6.6). For `hinf12`, SeDuMi solved all three problems accurately, and μ^* looks reasonable, whereas for `hinf13`, SeDuMi solved all three problems inaccurately, but μ^* still looks reasonable. Yet $c^T x_\delta^*$ is lower than $c^T x^*$ in both instances. We do not yet have an explanation for this.

The regularized solution x_δ^* has a one-norm lower than or equal to that of the unregularized solution x^* in all instances except `hinf12`, where $\|x_\delta^*\|_1$ is 1% higher (this small difference does not appear in Table 6.3). Solution sparsity is measured by

TABLE 6.3

Least one-norm solutions of the feasible DIMACS SDP/SOCPs. Three different types of SeDuMi failures are reported: [a]*numerical error;* [b]*primal infeasibility detected in solving* (6.7)*;* [c]*numerical error in solving* (6.7)*. The "schedule" instances have been abbreviated from* `sched_100_50_orig to sch_100_50_o`*, etc.*

SDP/SOCP	c^Tx^*	$c^Tx^*_\delta$	$\|x^*\|_1$	$\|x^*_\delta\|_1$	$\|x^*\|_0$	$\|x^*_\delta\|_0$	$\bar{\delta}$
nb	-5.07e−02	-5.07e−02	2.2e+0	↘ 2.1e+0	142 ↘	139	7.6e−3
nb_L1	-1.30e+01	-1.30e+01	3.1e+3	→ 3.1e+3	2407 ↘	1613	1.2e−5
nb_L2	-1.63e+00	-1.63e+00	3.1e+1	→ 3.1e+1	847 →	847	2.1e−5
nb_L2_bessel	-1.03e−01	-1.03e−01	1.0e+1	↘ 9.7e+0	131 ↗	133	2.7e−6
copo14	-3.11e−12	-2.13e−10	4.6e+0	↘ 2.0e+0	2128 ↘	224	4.7e−1
copo23	-8.38e−12	-3.73e−09	6.6e+0	↘ 2.0e+0	9430 ↘	575	4.7e−1
filter48_socp	1.42e+00	1.42e+00	7.6e+2	→ 7.6e+2	3284 ↘	3282	1.1e−6
minphase	5.98e+00	5.98e+00	1.6e+1	→ 1.6e+1	2304 →	2304	5.8e−2
hinf12	-3.68e−02	-7.11e−02	1.0e+0	→ 1.0e+0	138 ↗	194	5.1e+0
hinf13	-4.53e+01	-4.51e+01	2.8e+4	↘ 2.1e+4	322 ↘	318	2.8e−4
nql30	-9.46e−01	-9.46e−01	5.8e+3	↘ 2.8e+3	6301 →	6301	1.0e−7
nql60	-9.35e−01	-9.35e−01	2.3e+4	↘ 1.1e+4	25201 →	25201	1.4e−6
nql180	-9.28e−01	-9.28e−01	2.1e+5	↘ 1.0e+5	226776 ↘	226767	6.3e−8
nql30old	-9.46e−01	-9.46e−01	5.5e+3	↘ 1.0e+3	7502 ↘	6244	3.2e−5
nql60old	-9.35e−01	-9.35e−01	2.2e+4	↘ 4.0e+3	29515 ↘	23854	2.0e−5
nql180old	[a]-9.31e−01	[a]-9.29e−01	1.9e+5	↘ 6.8e+4	227097 ↘	211744	1.4e−8
qssp30	-6.50e+00	-6.50e+00	4.5e+3	→ 4.5e+3	7383 →	7383	4.1e−7
qssp60	-6.56e+00	-6.56e+00	1.8e+4	→ 1.8e+4	29163 →	29163	1.2e−6
qssp180	-6.64e+00	-6.64e+00	1.6e+5	→ 1.6e+5	260283 →	260283	[c]3.8e−7
sch_50_50_o	2.67e+04	2.67e+04	5.6e+4	→ 5.6e+4	1990 ↗	2697	8.7e−3
sch_50_50_s	7.85e+00	7.85e+00	1.1e+2	→ 1.1e+2	497 ↗	600	1.1e−5
sch_100_50_o	1.82e+05	1.82e+05	4.9e+5	→ 4.9e+5	3131 ↘	3040	2.4e−4
sch_100_50_s	6.72e+01	[b]8.69e+01	6.0e+4	↘ 1.3e+4	5827 ↗	7338	6.1e−3
sch_100_100_o	7.17e+05	[a]3.95e+02	1.8e+6	↘ 8.4e+2	12726 ↗	18240	1.3e−0
sch_100_100_s	2.73e+01	2.73e+01	1.6e+5	→ 1.6e+5	17574 ↘	16488	1.8e−8
sch_200_100_o	1.41e+05	1.41e+05	4.4e+5	→ 4.4e+5	24895 ↘	16561	4.3e−4
sch_200_100_s	5.18e+01	5.18e+01	7.8e+4	→ 7.8e+4	37271 ↘	37186	4.0e−8
truss5	1.33e+02	1.33e+02	2.1e+3	↘ 1.5e+3	3301 →	3301	1.6e−5
truss8	1.33e+02	1.33e+02	7.9e+3	↘ 5.2e+3	11914 ↘	11911	1.7e−7

the zero-norm defined in (6.4), where ϵ is based on the relative optimality gap

$$\epsilon = \frac{c^Tx^*_\delta - b^Ty^*_\delta}{1 + \|b\|\|y^*_\delta\| + \|c\|\|x^*_\delta\|}$$

of the computed solution of (6.6). For 52% of the instances, the regularized solution is sparser than the unregularized solution. For 28% of the instances, the solutions have the same sparsity. For the remaining six instances, the regularized solution is actually less sparse, even though its one-norm is lower (`nb_L2_bessel`, `sch_100_50_s`, `sch_100_100_o`) or the same (`hinf12`, `sch_50_50_o`, `sch_50_50_s`). SeDuMi implements an interior-point algorithm, so it is likely to find the analytic center of the solution set of (P).

The selection problem (6.7) is generally much harder to solve than (P) or (6.6). For example, on `nb_L2_bessel`, SeDuMi took 18, 99, and 16 iterations to solve (P), (6.7), and (6.6), respectively, and on `truss8` SeDuMi took, respectively, 24, 117, and 35 iterations. This seems to indicate that regularization is more efficient than solving the selection problem as a method for finding sparse solutions.

7. Discussion. We see from the numerical results in section 6 that regularization can provide an effective way of selecting a solution with desirable properties, such as

sparsity. However, finding the threshold value $\bar{\delta}$ for exact regularization entails first solving (P) to obtain p^*, and then solving (P^ϕ) to obtain μ^* and setting $\bar{\delta} = 1/\mu^*$; see Corollary 2.2. Can we find a $\delta < \bar{\delta}$ from (P) without also solving (P^ϕ)?

Consider the case of a CP, in which f and $\mathcal{C}$ have the form (1.1). Suppose that a value of $\delta < \bar{\delta}$ has been guessed (with $\bar{\delta}$ unknown), and a solution x^* of the regularized problem (P_δ) is obtained. By Corollary 2.2, x^* is also a solution of (P^ϕ). Suppose also that there exist Lagrange multipliers $y^* \in \mathbb{R}^m$ and $z^* \in \mathcal{K}^*$ for (P), where $\mathcal{K}^*$ is the dual cone of $\mathcal{K}$ given by

$$\mathcal{K}^* := \{y \in \mathbb{R}^n \mid y^T x \geq 0 \text{ for all } x \in \mathcal{K}\}.$$

Then (y^*, z^*) satisfy, among other conditions,

$$A^T y^* + z^* = c \qquad \text{and} \qquad b^T y^* = p^*.$$

Suppose, furthermore, that there exist Lagrange multipliers $y_\phi^* \in \mathbb{R}^m$, $z_\phi^* \in \mathcal{K}^*$, and $\mu^* \geq 0$ for (P^ϕ) that satisfy, among other conditions,

$$0 \in \partial\phi(x^*) - (A^T y_\phi^* + z_\phi^* - \mu^* c).$$

Then, analogous to the proof of Theorem 2.1, we can construct Lagrange multipliers for (P_δ) as follows:

Case 1. $\mu_\phi^* = 0$. The Lagrange multipliers for (P_δ) are given by

$$y_\delta^* := y^* + \delta y_\phi^* \quad \text{and} \quad z_\delta^* := z^* + \delta z_\phi^*.$$

Case 2. $\mu_\phi^* > 0$. The Lagrange multipliers for (P_δ) are given by

$$y_\delta^* := (1-\lambda) y^* + \frac{\lambda}{\mu_\phi^*} y_\phi^* \quad \text{and} \quad z_\delta^* := (1-\lambda) z^* + \frac{\lambda}{\mu_\phi^*} z_\phi^*,$$

for any $\lambda \in [0,1]$. The Lagrange multipliers (y_δ^*, z_δ^*) obtained for the regularized problem are therefore necessarily perturbed. Therefore, it is not possible to test the computed triple $(x^*, y_\delta^*, z_\delta^*)$ against the optimality conditions for the original CP in order to verify that x^* is indeed an exact solution.

In practice, if it were prohibitively expensive to solve (P) and (P^ϕ), we might adopt an approach suggested by Lucidi [27] and Mangasarian [33] for Tikhonov regularization. They suggest solving the regularized problem successively with decreasing values $\delta_1 > \delta_2 > \cdots$. If successive regularized solutions do not change for δ ranging over different orders of magnitude, then it is likely that a correct regularization parameter has been obtained. We note that in many instances, the threshold values $\bar{\delta}$ shown in Tables 6.1 and 6.2 are comfortably large, and a value such as $\delta = 10^{-4}$ would cover 85% of the these cases.

8. Appendix. In this appendix, we give an example of f and $\mathcal{C}$ that satisfy the assumptions of Theorem 5.2 and for which weak sharp minimum fails to hold and yet exact regularization holds for $\phi(x) = \|x - \widehat{x}\|_2^2$ and any $\widehat{x} \in \Re^n$.

Consider the example

$$n = 3, \quad f(x) = x_3, \quad \text{and} \quad \mathcal{C} = [0,1]^3 \cap \left(\cap_{k=2}^{\infty} \mathcal{C}^k\right),$$

where $\mathcal{C}^k = \{x \in \mathbb{R}^3 \mid x_1 - (k-1)x_2 - k^2 x_3 \leq 1/k\}$. Each $\mathcal{C}^k$ is a half-space in $\mathbb{R}^3$, so $\mathcal{C}$ is a closed convex set. Moreover, $\mathcal{C}$ is bounded and nonempty (since $0 \in \mathcal{C}$); see Figure 8.1(a). Clearly

$$p^* = 0 \quad \text{and} \quad \mathcal{S} = \{x \in \mathcal{C} \mid x_3 = 0\}. \tag{8.1}$$

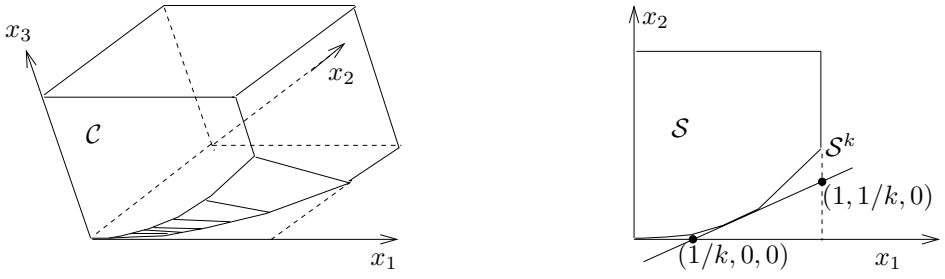


FIG. 8.1. (a) *The feasible set* $\mathcal{C}$. (b) *The solution set* $\mathcal{S}$ *and* $\mathcal{S}^k$.

First, we show that weak sharp minimum fails to hold, i.e., there does not exist $\tau > 0$ such that (5.1) holds with $\gamma = 1$. Let H^k be the hyperplane forming the boundary set of $\mathcal{C}^k$, i.e., $H^k = \{x \in \mathbb{R}^3 \mid x_1 - (k-1)x_2 - k^2 x_3 = 1/k\}$. Let x^k be the intersection point of H^k, H^{k+1} and the x_1x_3-plane. Direct calculation yields

$$x_1^k = \frac{1 - (1+1/k)^{-3}}{k(1-(1+1/k)^{-2})}, \quad x_2^k = 0, \quad x_3^k = \frac{x_1^k - 1/k}{k^2}, \tag{8.2}$$

for $k = 2, 3, \dots$. Since $\mathcal{C} \subset \mathcal{C}^k$, we have from (8.1) that $\mathcal{S} \subset \mathcal{S}^k$, where we let $\mathcal{S}^k = \{x \in \mathcal{C}^k \mid x_3 = 0\}$; see Figure 8.1(b). Thus

$$\operatorname{dist}(x^k, \mathcal{S}) \geq \operatorname{dist}(x^k, \mathcal{S}^k) \geq \operatorname{dist}((x_1^k, 0, 0), \mathcal{S}^k). \tag{8.3}$$

Since $\lim_{\alpha \to 1} \frac{1-\alpha^3}{1-\alpha^2} = \frac{3}{2}$, (8.2) implies that $kx_1^k \to 3/2$, i.e., $x_1^k = 1.5/k + o(1/k)$. The point in $\mathcal{S}^k$ nearest to $(x_1^k, 0, 0)$ lies on the line through $(1/k, 0, 0)$ and $(1, 1/k, 0)$ (with slope $1/(k-1)$ in the x_1x_2-plane), from which it follows that $\operatorname{dist}((x_1^k, 0, 0), \mathcal{S}^k) = 0.5/k^2 + o(1/k^2)$. Since $x_3^k = 1.5/k^3 + o(1/k^3)$ by (8.2), this together with (8.3) implies

$$\frac{x_3^k}{\operatorname{dist}(x^k, \mathcal{S})} \leq \frac{x_3^k}{\operatorname{dist}((x_1^k, 0, 0), \mathcal{S}^k)} = O(1/k) \to 0.$$

Moreover, for any $\ell \in \{2, 3, \dots\}$, we have from (8.2) and letting $\alpha = \ell/k$ that

$$\begin{aligned}
x_1^k - (\ell-1)x_2^k - \ell^2 x_3^k - \frac{1}{\ell} &= \left(1-\alpha^2\right) x_1^k - \frac{1}{\ell}\left(1-\alpha^3\right) \\
&= (1-\alpha)\left((1+\alpha)x_1^k - \frac{1}{\ell}\left(1+\alpha+\alpha^2\right)\right) \\
&= \frac{(1-\alpha)}{k}\left((1+\alpha)\frac{1-(1+1/k)^{-3}}{1-(1+1/k)^{-2}} - \frac{1}{\alpha}\left(1+\alpha+\alpha^2\right)\right) \\
&= \frac{(1-\alpha)}{k}(1+\alpha)\left(\frac{(1+1/k)^{-2}}{1+(1+1/k)^{-1}} - \frac{1}{\alpha(1+\alpha)}\right) \\
&= \frac{(1-\alpha^2)}{k}\left(\frac{k^2}{(2k+1)(k+1)} - \frac{k^2}{\ell(k+\ell)}\right) \\
&= \frac{(1-\alpha^2)}{k}\frac{(2k+\ell+1)(\ell-k-1)}{(2k+1)(k+1)\ell(k+\ell)},
\end{aligned}$$

where the second equality uses $1-\alpha^2 = (1-\alpha)(1+\alpha)$, $1-\alpha^3 = (1-\alpha)(1+\alpha+\alpha^2)$; the fourth equality uses the same identities but with $(1+1/k)^{-1}$ in place of α. By

considering the two cases $\ell \leq k$ and $\ell \geq k+1$, it is readily seen that the right-hand side of the previous equation is nonpositive. This in turn shows that $x^k \in \mathcal{C}^\ell$ for $\ell = 2, 3, \ldots$, and hence $x^k \in \mathcal{C}$.

Second, fix any $\widehat{x} \in \mathbb{R}^3$ and let $\phi(x) = \|x - \widehat{x}\|_2^2$. Let $x^* = \arg\min_{x \in \mathcal{S}} \phi(x)$ and $f_\delta(x) = f(x) + \delta\phi(x)$. Suppose $x^* \neq 0$. Then $\mathcal{C}$ is polyhedral in a neighborhood $\mathcal{N}$ of x^*. Since $x_\delta = \arg\min_{x \in \mathcal{C}} f_\delta(x)$ converges to x^* as $\delta \to 0$, we have that $x_\delta \in \mathcal{C} \cap \mathcal{N}$ for all $\delta > 0$ below some positive threshold, in which case exact regularization holds (see Corollary 2.3). Suppose $x^* = 0$. Then

$$\widehat{x} = -\nabla\phi(x^*) \in N_{\mathcal{S}}(x^*) = (-\infty, 0]^2 \times \mathbb{R},$$

where the second equality follows from $[0, \infty)^2 \times \{0\}$ being the tangent cone of $\mathcal{S}$ at 0. Thus $\widehat{x}_2 \leq 0, \widehat{x}_3 \leq 0$ and we see from

$$\nabla f_\delta(x^*) = (0, 0, 1)^T - \delta\widehat{x}$$

that $\nabla f_\delta(x^*) \geq 0$ for all $\delta \in [0, \bar{\delta}]$, where $\bar{\delta} = \infty$ if $\widehat{x}_3 \leq 0$ and $\bar{\delta} = 1/\widehat{x}_3$ if $\widehat{x}_3 > 0$. Because $\mathcal{C} \subset [0, \infty)^3$, it is implied that, for $\delta \in [0, \bar{\delta}]$,

$$\nabla f_\delta(x^*)^T(x - x^*) = \nabla f_\delta(x^*)^T x \geq 0 \quad \text{for all} \quad x \in \mathcal{C}.$$

Because $x^* \in \mathcal{C}$ and f_δ is strictly convex for $\delta \in (0, \bar{\delta}]$, it is implied that $x^* = \arg\min_{x \in \mathcal{C}} f_\delta(x)$ for all $\delta \in (0, \bar{\delta}]$. Hence exact regularization holds.

Acknowledgments. Sincere thanks to Dimitri Bertsekas for suggesting Lemma 4.1(a) and bringing to our attention the paper [4]. We also thank Kevin Leyton-Brown for generously giving us access to his CPLEX installation for the numerical experiments presented in section 6.

REFERENCES

[1] A. Altman and J. Gondzio, *Regularized symmetric indefinite systems in interior point methods for linear and quadratic optimization*, Optim. Methods Softw., 11 (1999), pp. 275–302.

[2] F. R. Bach, R. Thibaux, and M. I. Jordan, *Computing regularization paths for learning multiple kernels*, in Advances in Neural Information Processing Systems (NIPS) 17, L. Saul, Y. Weiss, and L. Bottou, eds., Morgan Kaufmann, San Mateo, CA, 2005.

[3] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*, MPS/SIAM Ser. Optim. 2, SIAM, Philadelphia, 2001.

[4] D. P. Bertsekas, *Necessary and sufficient conditions for a penalty method to be exact*, Math. Programming, 9 (1975), pp. 87–99.

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

[6] D. P. Bertsekas, *A note on error bounds for convex and nonconvex programs*, Comput. Optim. Appl., 12 (1999), pp. 41–51.

[7] D. P. Bertsekas, A. Nedić, and A. E. Ozdaglar, *Convex Analysis and Optimization*, Athena Scientific, Belmont, MA, 2003.

[8] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.

[9] J. V. Burke, *An exact penalization viewpoint of constrained optimization*, SIAM J. Control Optim., 29 (1991), pp. 968–998.

[10] J. V. Burke and S. Deng, *Weak sharp minima revisited.* II. *Application to linear regularity and error bounds*, Math. Program., 104 (2005), pp. 235–261.

[11] J. V. Burke and M. C. Ferris, *Weak sharp minima in mathematical programming*, SIAM J. Control Optim., 31 (1993), pp. 1340–1359.

[12] E. J. Candés, J. Romberg, and T. Tao, *Stable signal recovery from incomplete and inaccurate measurements*, Comm. Pure Appl. Math., 59 (2006), pp. 1207–1223.

[13] E. J. Candès, J. Romberg, and T. Tao, *Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information*, IEEE Trans. Inform. Theory, 52 (2006), pp. 489–509.
[14] S. S. Chen, D. L. Donoho, and M. A. Saunders, *Atomic decomposition by basis pursuit*, SIAM Rev., 43 (2001), pp. 129–159.
[15] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, *Trust-Region Methods*, MPS-SIAM Ser. Optim. 1, SIAM, Philadelphia, 2000.
[16] D. L. Donoho and M. Elad, *Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization*, Proc. Natl. Acad. Sci. USA, 100 (2003), pp. 2197–2202.
[17] D. L. Donoho, M. Elad, and V. Temlyakov, *Stable recovery of sparse overcomplete representations in the presence of noise*, IEEE Trans. Inform. Theory, 52 (2006), pp. 6–18.
[18] D. L. Donoho and J. Tanner, *Sparse nonnegative solution of underdetermined linear equations by linear programming*, Proc. Natl. Acad. Sci. USA, 102 (2005), pp. 9446–9451.
[19] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, *Least angle regression*, Ann. Statist., 32 (2004), pp. 407–499.
[20] M. C. Ferris and O. L. Mangasarian, *Finite perturbation of convex programs*, Appl. Math. Optim., 23 (1991), pp. 263–273.
[21] R. Fletcher, *An ℓ_1 penalty method for nonlinear constraints*, in Numerical Optimization, 1984, P. T. Boggs, R. H. Byrd, and R. B. Schnabel, eds., SIAM, Philadelphia, 1985, pp. 26–40.
[22] R. Fletcher, *Practical Methods of Optimization*, 2nd ed., John Wiley and Sons, Chichester, UK, 1987.
[23] C. C. Gonzaga, *Generation of Degenerate Linear Programming Problems*, Tech. report, Department of Mathematics, Federal University of Santa Catarina, Santa Catarina, Brazil, 2003.
[24] S.-P. Han and O. L. Mangasarian, *Exact penalty functions in nonlinear programming*, Math. Programming, 17 (1979), pp. 251–269.
[25] C. Kanzow, H. Qi, and L. Qi, *On the minimum norm solution of linear programs*, J. Optim. Theory Appl., 116 (2003), pp. 333–345.
[26] S. Lucidi, *A finite algorithm for the least two-norm solution of a linear program*, Optimization, 18 (1987), pp. 809–823.
[27] S. Lucidi, *A new result in the theory and computation of the least-norm solution of a linear program*, J. Optim. Theory Appl., 55 (1987), pp. 103–117.
[28] Z.-Q. Luo and J.-S. Pang, *Error bounds for analytic systems and their applications*, Math. Programming, 67 (1994), pp. 1–28.
[29] Z.-Q. Luo and J.-S. Pang, eds., *Error bounds in mathematical programming*, Math. Program., 88 (2000), pp. 221–410.
[30] O. L. Mangasarian, *Normal solutions of linear programs*, Math. Programming Stud., 22 (1984), pp. 206–216.
[31] O. L. Mangasarian, *Sufficiency of exact penalty minimization*, SIAM J. Control Optim., 23 (1985), pp. 30–37.
[32] O. L. Mangasarian, *A simple characterization of solution sets of convex programs*, Oper. Res. Lett., 7 (1988), pp. 21–26.
[33] O. L. Mangasarian, *A Newton method for linear programming*, J. Optim. Theory Appl., 121 (2004), pp. 1–18.
[34] O. L. Mangasarian and R. R. Meyer, *Nonlinear perturbation of linear programs*, SIAM J. Control Optim., 17 (1979), pp. 745–752.
[35] Y. E. Nesterov and A. Nemirovski, *Interior Point Polynomial Algorithms in Convex Programming*, SIAM Stud. Appl. Math. 13, SIAM, Philadelphia, 1994.
[36] *NETLIB Linear Programming Library*, available online at http://www.netlib.org/lp/infeas/, 2006.
[37] G. Pataki and S. Schmieta, *The DIMACS Library of Semidefinite-Quadratic-Linear Programs*, Tech. report preliminary draft, Computational Optimization Research Center, Columbia University, New York, 2002.
[38] J. Renegar, *A Mathematical View of Interior-Point Methods in Convex Optimization*, MPS/SIAM Ser. Optim. 3, SIAM, Philadelphia, 2001.
[39] S. M. Robinson, *Local structure of feasible sets in nonlinear programming.* II. *Nondegeneracy*, Math. Programming Stud., 22 (1984), pp. 217–230.
[40] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[41] S. Sardy and P. Tseng, *On the statistical analysis of smoothing by maximizing dirty Markov random field posterior distributions*, J. Amer. Statist. Assoc., 99 (2004), pp. 191–204.
[42] M. A. Saunders, *Cholesky-based methods for sparse least squares: The benefits of regular-*

ization, in Linear and Nonlinear Conjugate Gradient-Related Methods, L. Adams and J. L. Nazareth, eds., SIAM, Philadelphia, 1996, pp. 92–100.

[43] J. F. Sturm, *Using Sedumi* 1.02*, a Matlab Toolbox for Optimization over Symmetric Cones (updated for Version* 1.05*)*, Tech. report, Department of Econometrics, Tilburg University, Tilburg, The Netherlands, 2001.

[44] R. Tibshirani, *Regression shrinkage and selection via the Lasso*, J. Roy. Statist. Soc. Ser. B, 58 (1996), pp. 267–288.

[45] A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-Posed Problems*, V. H. Winston and Sons, Washington, DC, 1977 (translated from Russian).

[46] Z. Wu and J. J. Ye, *On error bounds for lower semicontinuous functions*, Math. Program., 92 (2002), pp. 301–314.

[47] Y. Ye, *Interior-Point Algorithms: Theory and Analysis*, John Wiley and Sons, New York, 1997.

[48] Y.-B. Zhao and D. Li, *Locating the least* 2*-norm solution of linear programs via a path-following method*, SIAM J. Optim., 12 (2002), pp. 893–912.

SIAM J. OPTIM.
Vol. 18, No. 4, pp. 1351–1376

PROXIMAL THRESHOLDING ALGORITHM FOR MINIMIZATION OVER ORTHONORMAL BASES*

PATRICK L. COMBETTES† AND JEAN-CHRISTOPHE PESQUET‡

Abstract. The notion of soft thresholding plays a central role in problems from various areas of applied mathematics, in which the ideal solution is known to possess a sparse decomposition in some orthonormal basis. Using convex-analytical tools, we extend this notion to that of proximal thresholding and investigate its properties, providing, in particular, several characterizations of such thresholders. We then propose a versatile convex variational formulation for optimization over orthonormal bases that covers a wide range of problems, and we establish the strong convergence of a proximal thresholding algorithm to solve it. Numerical applications to signal recovery are demonstrated.

Key words. convex programming, deconvolution, denoising, forward-backward splitting algorithm, Hilbert space, orthonormal basis, proximal algorithm, proximal thresholding, proximity operator, signal recovery, soft thresholding, strong convergence

AMS subject classifications. 90C25, 65K10, 94A12

DOI. 10.1137/060669498

1. Problem formulation.

Throughout this paper, $\mathcal{H}$ is a separable infinite-dimensional real Hilbert space with scalar product $\langle \cdot \mid \cdot \rangle$, norm $\| \cdot \|$, and distance d. Moreover, $\Gamma_0(\mathcal{H})$ denotes the class of proper lower semicontinuous convex functions from $\mathcal{H}$ to $]-\infty,+\infty]$, and $(e_k)_{k\in\mathbb{N}}$ is an orthonormal basis of $\mathcal{H}$.

The standard denoising problem in signal theory consists of recovering the original form of a signal $\overline{x} \in \mathcal{H}$ from an observation $z = \overline{x} + v$, where $v \in \mathcal{H}$ is the realization of a noise process. In many instances, $\overline{x}$ is known to admit a sparse representation with respect to $(e_k)_{k\in\mathbb{N}}$, and an estimate x of $\overline{x}$ can be constructed by removing the coefficients of small magnitude in the representation $(\langle z \mid e_k \rangle)_{k\in\mathbb{N}}$ of z with respect to $(e_k)_{k\in\mathbb{N}}$. A popular method consists of performing a so-called soft thresholding of each coefficient $\langle z \mid e_k \rangle$ at some predetermined level $\omega_k \in]0,+\infty[$, namely

$$x = \sum_{k\in\mathbb{N}} \mathrm{soft}_{[-\omega_k,\omega_k]}\left(\langle z \mid e_k \rangle\right) e_k, \tag{1.1}$$

where (see Figure 2.1)

$$\mathrm{soft}_{[-\omega_k,\omega_k]} : \xi \mapsto \mathrm{sign}(\xi)\max\{|\xi| - \omega_k, 0\}. \tag{1.2}$$

This approach has received considerable attention in various areas of applied mathematics ranging from nonlinear approximation theory to statistics, and from harmonic analysis to image processing; see, for instance, [2, 7, 9, 21, 23, 29, 33] and the references therein. From an optimization point of view (see Remark 2.8), the vector x

*Received by the editors September 10, 2006; accepted for publication (in revised form) April 25, 2007; published electronically November 14, 2007.
http://www.siam.org/journals/siopt/18-4/66949.html
†Laboratoire Jacques-Louis Lions, UMR CNRS 7598, Faculté de Mathématiques, Université Pierre et Marie Curie — Paris 6, 75005 Paris, France (plc@math.jussieu.fr).
‡Institut Gaspard Monge and UMR CNRS 8049, Université de Marne la Vallée, 77454 Marne la Vallée Cedex 2, France (pesquet@univ-mlv.fr).

exhibited in (1.1) is the solution to the variational problem

$$\underset{x\in\mathcal{H}}{\text{minimize}}\ \ \frac{1}{2}\|x-z\|^2+\sum_{k\in\mathbb{N}}\omega_k\,|\langle x\mid e_k\rangle|\,. \tag{1.3}$$

Attempts have been made to extend this formulation to the more general inverse problems in which the observation assumes the form $z=T\overline{x}+v$, where T is a nonzero bounded linear operator from $\mathcal{H}$ to some real Hilbert space $\mathcal{G}$, and where $v\in\mathcal{G}$ is the realization of a noise process. Thus, the variational problem

$$\underset{x\in\mathcal{H}}{\text{minimize}}\ \ \frac{1}{2}\|Tx-z\|^2+\sum_{k\in\mathbb{N}}\omega_k\,|\langle x\mid e_k\rangle| \tag{1.4}$$

has been considered and, since it admits no closed-form solution, the soft thresholding algorithm

$$x_0\in\mathcal{H}\quad\text{and}\quad(\forall n\in\mathbb{N})\ \ x_{n+1}=\sum_{k\in\mathbb{N}}\operatorname{soft}_{[-\omega_k,\omega_k]}\big(\langle x_n+T^*(z-Tx_n)\mid e_k\rangle\big)e_k \tag{1.5}$$

has been proposed to solve it [5, 19, 20, 24] (see also [36] and the references therein for related work). The strong convergence of this algorithm was formally established in [18].

PROPOSITION 1.1 (see [18, Theorem 3.1]). *Suppose that* $\inf_{k\in\mathbb{N}}\omega_k>0$ *and* $\|T\|<1$. *Then the sequence* $(x_n)_{n\in\mathbb{N}}$ *generated by* (1.5) *converges strongly to a solution to* (1.4).

In [16], (1.4) was analyzed in a broader framework, and the following extension of Proposition 1.1 was obtained by bringing into play tools from convex analysis and recent results from constructive fixed point theory (Proposition 1.2 reduces to Proposition 1.1 when $\|T\|<1$, $\gamma_n\equiv 1$, and $\lambda_n\equiv 1$).

PROPOSITION 1.2 (see [16, Corollary 5.19]). *Let* $(\gamma_n)_{n\in\mathbb{N}}$ *be a sequence in* $]0,+\infty[$, *and let* $(\lambda_n)_{n\in\mathbb{N}}$ *be a sequence in* $]0,1]$. *Suppose that the following hold:* $\inf_{k\in\mathbb{N}}\omega_k>0$, $\inf_{n\in\mathbb{N}}\gamma_n>0$, $\sup_{n\in\mathbb{N}}\gamma_n<2/\|T\|^2$, *and* $\inf_{n\in\mathbb{N}}\lambda_n>0$. *Then the sequence* $(x_n)_{n\in\mathbb{N}}$ *generated by the algorithm*

$$\begin{aligned}x_0\in\mathcal{H}\quad\text{and}\quad(\forall n\in\mathbb{N})\quad x_{n+1}&=x_n\\&+\lambda_n\Bigg(\sum_{k\in\mathbb{N}}\operatorname{soft}_{[-\gamma_n\omega_k,\gamma_n\omega_k]}\big(\langle x_n+\gamma_nT^*(z-Tx_n)\mid e_k\rangle\big)e_k-x_n\Bigg)\end{aligned} \tag{1.6}$$

converges strongly to a solution to (1.4).

In denoising and approximation problems, various theoretical, physical, and heuristic considerations have led researchers to consider alternative thresholding strategies in (1.1); see, e.g., [1, 33, 34, 35, 39]. However, the questions of whether alternative thresholding rules can be used in algorithms akin to (1.6) and of identifying the underlying variational problems remain open. These questions are significant because the current theory of iterative thresholding, as described by Proposition 1.2, can tackle only variational problems of the form (1.4), which offers limited flexibility in the penalization of the coefficients $(\langle x\mid e_k\rangle)_{k\in\mathbb{N}}$ and which is furthermore restricted to standard linear inverse problems. The aim of the present paper is to bring out general answers to these questions. Our analysis will revolve around the following variational formulation, where σ_Ω denotes the support function of a set Ω (see (2.2)).

PROBLEM 1.3. Let $\Phi \in \Gamma_0(\mathcal{H})$, let $\mathbb{K} \subset \mathbb{N}$, let $\mathbb{L} = \mathbb{N} \smallsetminus \mathbb{K}$, let $(\Omega_k)_{k\in\mathbb{K}}$ be a sequence of closed intervals in $\mathbb{R}$, and let $(\psi_k)_{k\in\mathbb{N}}$ be a sequence in $\Gamma_0(\mathbb{R})$. The objective is to

$$\underset{x\in\mathcal{H}}{\text{minimize}} \;\; \Phi(x) + \sum_{k\in\mathbb{N}} \psi_k(\langle x \mid e_k\rangle) + \sum_{k\in\mathbb{K}} \sigma_{\Omega_k}(\langle x \mid e_k\rangle), \tag{1.7}$$

under the following standing assumptions:

(i) the function Φ is differentiable on $\mathcal{H}$, $\inf \Phi(\mathcal{H}) > -\infty$, and $\nabla\Phi$ is $1/\beta$-Lipschitz continuous for some $\beta \in \left]0, +\infty\right[$;
(ii) for every $k \in \mathbb{N}$, $\psi_k \geq \psi_k(0) = 0$;
(iii) the functions $(\psi_k)_{k\in\mathbb{N}}$ are differentiable at 0;
(iv) if $\mathbb{L} \neq \varnothing$, then the functions $(\psi_k)_{k\in\mathbb{L}}$ are finite and twice differentiable on $\mathbb{R} \smallsetminus \{0\}$, and

$$(\forall \rho \in \left]0, +\infty\right[)(\exists\, \theta \in \left]0, +\infty\right[) \quad \inf_{k\in\mathbb{L}} \inf_{0<|\xi|\leq\rho} \psi_k''(\xi) \geq \theta; \tag{1.8}$$

(v) if $\mathbb{L} \neq \varnothing$, then the function $\Upsilon_{\mathbb{L}} \colon \ell^2(\mathbb{L}) \to \left]-\infty, +\infty\right] \colon (\xi_k)_{k\in\mathbb{L}} \mapsto \sum_{k\in\mathbb{L}} \psi_k(\xi_k)$ is coercive;
(vi) $(\exists\, \omega \in \left]0, +\infty\right[)\ [-\omega, \omega] \subset \bigcap_{k\in\mathbb{K}} \Omega_k$.

Let us note that Problem 1.3 reduces to (1.4) when $\Phi \colon x \mapsto \|Tx - z\|^2/2$, $\mathbb{K} = \mathbb{N}$, and, for every $k \in \mathbb{N}$, $\Omega_k = [-\omega_k, \omega_k]$, and $\psi_k = 0$. It will be shown (Proposition 4.1) that Problem 1.3 admits at least one solution. While assumption (i) on Φ may seem offhand to be rather restrictive, it will be seen in section 5.1 to cover important scenarios. In addition, it makes it possible to employ a forward-backward splitting strategy to solve (1.7), which consists essentially of alternating a forward (explicit) gradient step on Φ with a backward (implicit) proximal step on

$$\Psi \colon \mathcal{H} \to \left]-\infty, +\infty\right] \colon x \mapsto \sum_{k\in\mathbb{N}} \psi_k(\langle x \mid e_k\rangle) + \sum_{k\in\mathbb{K}} \sigma_{\Omega_k}(\langle x \mid e_k\rangle). \tag{1.9}$$

Our main convergence result (Theorem 4.5) will establish the *strong* convergence of an inexact forward-backward splitting algorithm (Algorithm 4.3) for solving Problem 1.3. Another contribution of this paper will be to show (Remark 3.4) that, under our standing assumptions, the function displayed in (1.9) is quite general in the sense that the operators on $\mathcal{H}$ that perform nonexpansive (as required by our convergence analysis) and increasing (as imposed by practical considerations) thresholdings on the closed intervals $(\Omega_k)_{k\in\mathbb{K}}$ of the coefficients $(\langle x \mid e_k\rangle)_{k\in\mathbb{K}}$ of a point $x \in \mathcal{H}$ are precisely those of the form prox_Ψ, i.e., the proximity operator of Ψ. Furthermore, we show (Proposition 3.6 and Lemma 2.3) that such an operator, which provides the proximal step of our algorithm, can be conveniently decomposed as

$$\operatorname{prox}_\Psi \colon \mathcal{H} \to \mathcal{H} \colon x \mapsto \sum_{k\in\mathbb{K}} \operatorname{prox}_{\psi_k}\big(\operatorname{soft}_{\Omega_k}\langle x \mid e_k\rangle\big) e_k + \sum_{k\in\mathbb{L}} \operatorname{prox}_{\psi_k}\langle x \mid e_k\rangle\, e_k, \tag{1.10}$$

where we define the soft thresholder relative to a nonempty closed interval $\Omega \subset \mathbb{R}$ as

$$\operatorname{soft}_\Omega : \mathbb{R} \to \mathbb{R} \colon \xi \mapsto \begin{cases} \xi - \underline{\omega} & \text{if } \xi < \underline{\omega}, \\ 0 & \text{if } \xi \in \Omega, \\ \xi - \overline{\omega} & \text{if } \xi > \overline{\omega}, \end{cases} \qquad \text{with} \quad \begin{cases} \underline{\omega} = \inf \Omega, \\ \overline{\omega} = \sup \Omega. \end{cases} \tag{1.11}$$

The remainder of the paper is organized as follows. In section 2, we provide a brief account of the theory of proximity operators, which play a central role in our analysis. In section 3, we introduce and study the notion of a proximal thresholder. Our algorithm is presented in section 4 and its strong convergence to a solution to Problem 1.3 is demonstrated. Signal recovery applications are discussed in section 5, where numerical results are presented.

2. Proximity operators. Let us first introduce some basic notation (for a detailed account of convex analysis, see [41]). Let C be a subset of $\mathcal{H}$. The indicator function of C is

$$\iota_C\colon \mathcal{H} \to \{0,+\infty\}\colon x \mapsto \begin{cases} 0 & \text{if } x \in C,\\ +\infty & \text{if } x \notin C, \end{cases} \tag{2.1}$$

its support function is

$$\sigma_C\colon \mathcal{H} \to [-\infty,+\infty] : u \mapsto \sup_{x\in C} \langle x \mid u\rangle, \tag{2.2}$$

and its distance function is $d_C\colon \mathcal{H} \to [0,+\infty] : x \mapsto \inf \|C - x\|$. If C is nonempty, closed, and convex, then, for every $x \in \mathcal{H}$, there exists a unique point $P_C x \in C$, called the projection of x onto C, such that $\|x - P_C x\| = d_C(x)$. A function $f\colon \mathcal{H} \to [-\infty,+\infty]$ is proper if $-\infty \notin f(\mathcal{H}) \neq \{+\infty\}$, and coercive if $\lim_{\|x\|\to+\infty} f(x) = +\infty$. The domain of $f\colon \mathcal{H} \to [-\infty,+\infty]$ is $\operatorname{dom} f = \big\{x \in \mathcal{H} \;\big|\; f(x) < +\infty\big\}$, its set of global minimizers is denoted by $\operatorname{Argmin} f$, and its conjugate is the function $f^*\colon \mathcal{H} \to [-\infty,+\infty] : u \mapsto \sup_{x\in\mathcal{H}} \langle x \mid u\rangle - f(x)$; if f is proper, its subdifferential is the set-valued operator

$$\partial f\colon \mathcal{H} \to 2^{\mathcal{H}}\colon x \mapsto \big\{u \in \mathcal{H} \;\big|\; (\forall y \in \operatorname{dom} f)\ \langle y - x \mid u\rangle + f(x) \leq f(y)\big\}. \tag{2.3}$$

If $f\colon \mathcal{H} \to \left]-\infty,+\infty\right]$ is convex and Gâteaux differentiable at $x \in \operatorname{dom} f$ with gradient $\nabla f(x)$, then $\partial f(x) = \{\nabla f(x)\}$.

Example 2.1. Let $\Omega \subset \mathbb{R}$ be a nonempty closed interval, let $\underline{\omega} = \inf \Omega$, let $\overline{\omega} = \sup \Omega$, and let $\xi \in \mathbb{R}$. Then the following hold:

(i) $$\sigma_\Omega(\xi) = \begin{cases} \underline{\omega}\xi & \text{if } \xi < 0,\\ 0 & \text{if } \xi = 0,\\ \overline{\omega}\xi & \text{if } \xi > 0. \end{cases}$$

(ii) $$\partial\sigma_\Omega(\xi) = \begin{cases} \{\underline{\omega}\} \cap \mathbb{R} & \text{if } \xi < 0,\\ \Omega & \text{if } \xi = 0,\\ \{\overline{\omega}\} \cap \mathbb{R} & \text{if } \xi > 0. \end{cases}$$

The infimal convolution of two functions $f, g\colon \mathcal{H} \to \left]-\infty,+\infty\right]$ is denoted by $f \,\square\, g$. Finally, an operator $R\colon \mathcal{H} \to \mathcal{H}$ is nonexpansive if $(\forall (x,y) \in \mathcal{H}^2)\ \|Rx - Ry\| \leq \|x - y\|$ and firmly nonexpansive if $(\forall (x,y) \in \mathcal{H}^2)\ \|Rx - Ry\|^2 \leq \langle x - y \mid Rx - Ry\rangle$.

Proximity operators (sometimes called "proximal mappings") were introduced by Moreau [30] and their use in signal theory goes back to [11] (see also [8, 16] for recent developments). We briefly recall some essential facts below and refer the reader to [16] and [31] for more details. Let $f \in \Gamma_0(\mathcal{H})$. The proximity operator of f is the operator $\operatorname{prox}_f\colon \mathcal{H} \to \mathcal{H}$ which maps every $x \in \mathcal{H}$ to the unique minimizer of the function $y \mapsto f(y) + \|x - y\|^2/2$. It is characterized by

$$(\forall x \in \mathcal{H})(\forall p \in \mathcal{H}) \quad p = \operatorname{prox}_f x \quad \Leftrightarrow \quad x - p \in \partial f(p). \tag{2.4}$$

LEMMA 2.2. *Let $f \in \Gamma_0(\mathcal{H})$. Then the following hold:*
(i) $(\forall x \in \mathcal{H})\ \big[\, x \in \operatorname{Argmin} f \Leftrightarrow 0 \in \partial f(x) \Leftrightarrow \operatorname{prox}_f x = x \,\big]$.
(ii) $\operatorname{prox}_{f^*} = \operatorname{Id} - \operatorname{prox}_f$.
(iii) prox_f *is firmly nonexpansive.*
(iv) *If f is even, then* prox_f *is odd.*

The next result provides a key decomposition property with respect to the orthonormal basis $(e_k)_{k\in\mathbb{N}}$.

LEMMA 2.3 (see [16, Example 2.19]). *Set*

$$f\colon \mathcal{H} \to \left]-\infty,+\infty\right] : x \mapsto \sum_{k\in\mathbb{N}} \phi_k(\langle x \mid e_k\rangle), \tag{2.5}$$

where $(\phi_k)_{k\in\mathbb{N}}$ are functions in $\Gamma_0(\mathbb{R})$ that satisfy $(\forall k \in \mathbb{N})\ \phi_k \geq \phi_k(0) = 0$. Then $f \in \Gamma_0(\mathcal{H})$ and $(\forall x \in \mathcal{H})\ \operatorname{prox}_f x = \sum_{k\in\mathbb{N}} \operatorname{prox}_{\phi_k}\langle x \mid e_k\rangle\, e_k$.

The remainder of this section is dedicated to proximity operators on the real line, the importance of which is underscored by Lemma 2.3.

PROPOSITION 2.4. *Let ϱ be a function defined from $\mathbb{R}$ to $\mathbb{R}$. Then ϱ is the proximity operator of a function in $\Gamma_0(\mathbb{R})$ if and only if it is nonexpansive and increasing.*

Proof. Let ξ and η be real numbers. First, suppose that $\varrho = \operatorname{prox}_\phi$, where $\phi \in \Gamma_0(\mathbb{R})$. Then it follows from Lemma 2.2(iii) that ϱ is nonexpansive and that $0 \leq |\varrho(\xi) - \varrho(\eta)|^2 \leq (\xi-\eta)(\varrho(\xi)-\varrho(\eta))$, which shows that ϱ is increasing since $\xi-\eta$ and $\varrho(\xi)-\varrho(\eta)$ have the same sign. Conversely, suppose that ϱ is nonexpansive and increasing and, without loss of generality, that $\xi \leq \eta$. Then, $0 \leq \varrho(\xi)-\varrho(\eta) \leq \xi-\eta$ and therefore $|\varrho(\xi)-\varrho(\eta)|^2 \leq (\xi-\eta)(\varrho(\xi)-\varrho(\eta))$. Thus, ϱ is firmly nonexpansive. However, every firmly nonexpansive operator $R\colon \mathcal{H} \to \mathcal{H}$ is of the form $R = (\operatorname{Id} + A)^{-1}$, where $A\colon \mathcal{H} \to 2^{\mathcal{H}}$ is a maximal monotone operator [6]. Since the only maximal monotone operators in $\mathbb{R}$ are subdifferentials of functions in $\Gamma_0(\mathbb{R})$ [32, section 24], we must have $\varrho = (\operatorname{Id} + \partial\phi)^{-1} = \operatorname{prox}_\phi$ for some $\phi \in \Gamma_0(\mathbb{R})$. ∎

COROLLARY 2.5. *Suppose that 0 is a minimizer of $\phi \in \Gamma_0(\mathbb{R})$. Then*

$$(\forall \xi \in \mathbb{R}) \quad \begin{cases} 0 \leq \operatorname{prox}_\phi \xi \leq \xi & \text{if } \xi > 0, \\ \operatorname{prox}_\phi \xi = 0 & \text{if } \xi = 0, \\ \xi \leq \operatorname{prox}_\phi \xi \leq 0 & \text{if } \xi < 0. \end{cases} \tag{2.6}$$

This is true, in particular, when ϕ is even, in which case prox_ϕ *is an odd operator.*

Proof. Since $0 \in \operatorname{Argmin}\phi$, Lemma 2.2(i) yields $\operatorname{prox}_\phi 0 = 0$. In turn, since prox_ϕ is nonexpansive by Lemma 2.2(iii), we have $(\forall \xi \in \mathbb{R})\ |\operatorname{prox}_\phi \xi| = |\operatorname{prox}_\phi \xi - \operatorname{prox}_\phi 0| \leq |\xi - 0| = |\xi|$. Altogether, since Proposition 2.4 asserts that prox_ϕ is increasing, we obtain (2.6). Finally, if ϕ is even, its convexity yields $(\forall \xi \in \operatorname{dom}\phi)$ $\phi(0) = \phi\big((\xi-\xi)/2\big) \leq \big(\phi(\xi)+\phi(-\xi)\big)/2 = \phi(\xi)$. Therefore $0 \in \operatorname{Argmin}\phi$, while the oddness of prox_ϕ follows from Lemma 2.2(iv). ∎

Let us now provide some elementary examples (Example 2.6 is illustrated in Figure 2.1 in the case when $\Omega = [-1,1]$).

Example 2.6. Let $\Omega \subset \mathbb{R}$ be a nonempty closed interval, let $\underline{\omega} = \inf\Omega$, let $\overline{\omega} = \sup\Omega$, and let $\xi \in \mathbb{R}$. Then the following hold:

(i) $\operatorname{prox}_{\iota_\Omega} \xi = P_\Omega \xi = \begin{cases} \underline{\omega} & \text{if } \xi < \underline{\omega}, \\ \xi & \text{if } \xi \in \Omega, \\ \overline{\omega} & \text{if } \xi > \overline{\omega}. \end{cases}$

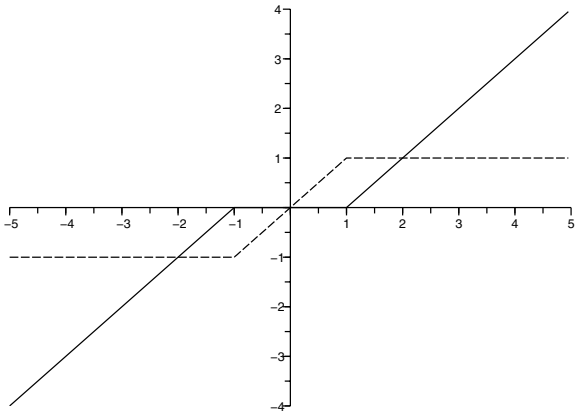


FIG. 2.1. *Graphs of* $\operatorname{prox}_{\phi} = \operatorname{soft}_{[-1,1]}$ *(solid line) and* $\operatorname{prox}_{\phi^*} = P_{[-1,1]}$ *(dashed line), where* $\phi = |\cdot|$ *and* $\phi^* = \iota_{[-1,1]}$ *(see Example* 2.6*).*

(ii) $\operatorname{prox}_{\sigma_\Omega}\xi = \operatorname{soft}_\Omega \xi$, where $\operatorname{soft}_\Omega$ is the soft thresholder defined in (1.11).

Proof. (i) is clear and, since $\sigma_\Omega^* = \iota_\Omega$, (ii) follows from (i) and Lemma 2.2(ii). □

Example 2.7 (see [8, Examples 4.2 and 4.4]). Let $p \in [1,+\infty[$, let $\omega \in]0,+\infty[$, let $\phi\colon \mathbb{R} \to \mathbb{R}\colon \eta \mapsto \omega|\eta|^p$, let $\xi \in \mathbb{R}$, and set $\pi = \operatorname{prox}_\phi \xi$. Then the following hold:

(i) $\pi = \operatorname{soft}_{[-\omega,\omega]}(\xi) = \operatorname{sign}(\xi)\max\{|\xi| - \omega, 0\}$ if $p = 1$;

(ii) $\pi = \xi + \dfrac{4\omega}{3\cdot 2^{1/3}}((\rho-\xi)^{1/3} - (\rho+\xi)^{1/3})$, where $\rho = \sqrt{\xi^2 + 256\omega^3/729}$ if $p = 4/3$;

(iii) $\pi = \xi + 9\omega^2 \operatorname{sign}(\xi)\big(1 - \sqrt{1 + 16|\xi|/(9\omega^2)}\,\big)/8$ if $p = 3/2$;

(iv) $\pi = \xi/(1+2\omega)$ if $p = 2$;

(v) $\pi = \operatorname{sign}(\xi)\big(\sqrt{1+12\omega|\xi|} - 1\big)/(6\omega)$ if $p = 3$;

(vi) $\pi = \big(\frac{\rho+\xi}{8\omega}\big)^{1/3} - \big(\frac{\rho-\xi}{8\omega}\big)^{1/3}$, where $\rho = \sqrt{\xi^2 + 1/(27\omega)}$ if $p = 4$.

Remark 2.8. The variational problem described in (1.3) is equivalent to minimizing over $\mathcal{H}$ the function $x \mapsto f(x) + \|z - x\|^2/2$, where $f\colon \mathcal{H} \to]-\infty,+\infty] : x \mapsto \sum_{k\in\mathbb{N}} \omega_k |\langle x \mid e_k\rangle|$. In view of Lemma 2.3 and Example 2.7(i), its solution is $\operatorname{prox}_f z = \sum_{k\in\mathbb{N}} \operatorname{soft}_{[-\omega_k,\omega_k]}(\langle z \mid e_k\rangle)e_k$, as displayed in (1.1).

PROPOSITION 2.9. *Let* ψ *be a function in* $\Gamma_0(\mathbb{R})$*, and let* ρ *and* θ *be real numbers in* $]0,+\infty[$ *such that*

(i) $\psi \geq \psi(0) = 0$,

(ii) ψ *is differentiable at* 0,

(iii) ψ *is twice differentiable on* $[-\rho,\rho] \smallsetminus \{0\}$ *and* $\inf_{0<|\xi|\leq\rho} \psi''(\xi) \geq \theta$.

Then $(\forall \xi \in [-\rho,\rho])(\forall \eta \in [-\rho,\rho])\ |\operatorname{prox}_\psi \xi - \operatorname{prox}_\psi \eta| \leq |\xi - \eta|/(1+\theta)$.

Proof. Set $R = [-\rho,\rho] \smallsetminus \{0\}$ and $\varphi\colon R \to \mathbb{R}\colon \zeta \mapsto \zeta + \psi'(\zeta)$. We first infer from (iii) that

$$(\forall \zeta \in R) \quad \varphi'(\zeta) = 1 + \psi''(\zeta) \geq 1 + \theta. \tag{2.7}$$

Moreover, (2.4) yields $(\forall \zeta \in R)\ \operatorname{prox}_\psi \zeta = \varphi^{-1}(\zeta)$. Note also that, in light of (2.4), (ii), and (i), we have $(\forall \zeta \in \mathbb{R})\ \operatorname{prox}_\psi \zeta = 0 \Leftrightarrow \zeta \in \partial\psi(0) = \{\psi'(0)\} = \{0\}$. Hence,

prox_ψ vanishes only at 0 and we derive from Lemma 2.2(iii) that

$$(\forall \zeta \in R) \quad 0 < |\varphi^{-1}(\zeta)| = |\operatorname{prox}_\psi \zeta - \operatorname{prox}_\psi 0| \leq |\zeta - 0| \leq \rho. \tag{2.8}$$

In turn, we deduce from (2.7) that

$$\sup_{\zeta \in R} \operatorname{prox}'_\psi \zeta = \frac{1}{\inf_{\zeta \in R} \varphi'\big(\varphi^{-1}(\zeta)\big)} \leq \frac{1}{\inf_{\zeta \in R} \varphi'(\zeta)} \leq \frac{1}{1+\theta}. \tag{2.9}$$

Now fix ξ and η in R. First, let us assume that either $\xi < \eta < 0$ or $0 < \xi < \eta$. Then, since prox_ψ is increasing by Proposition 2.4, it follows from the mean value theorem and (2.9) that there exists $\mu \in]\xi, \eta[$ such that

$$0 \leq \operatorname{prox}_\psi \eta - \operatorname{prox}_\psi \xi = (\eta - \xi) \operatorname{prox}'_\psi \mu \leq (\eta - \xi) \sup_{\zeta \in R} \operatorname{prox}'_\psi \zeta \leq \frac{\eta - \xi}{1+\theta}. \tag{2.10}$$

Next, let us assume that $\xi < 0 < \eta$. Then the mean value theorem asserts that there exist $\mu \in]\xi, 0[$ and $\nu \in]0, \eta[$ such that

$$\operatorname{prox}_\psi 0 - \operatorname{prox}_\psi \xi = -\xi \operatorname{prox}'_\psi \mu \quad \text{and} \quad \operatorname{prox}_\psi \eta - \operatorname{prox}_\psi 0 = \eta \operatorname{prox}'_\psi \nu. \tag{2.11}$$

Since prox_ψ is increasing and $\operatorname{prox}_\psi 0 = 0$, we obtain

$$0 \leq \operatorname{prox}_\psi \eta - \operatorname{prox}_\psi \xi = \eta \operatorname{prox}'_\psi \nu - \xi \operatorname{prox}'_\psi \mu \leq (\eta - \xi) \sup_{\zeta \in R} \operatorname{prox}'_\psi \zeta \leq \frac{\eta - \xi}{1+\theta}. \tag{2.12}$$

Altogether, we have shown that, for every ξ and η in R, $|\operatorname{prox}_\psi \xi - \operatorname{prox}_\psi \eta| \leq |\xi - \eta|/(1+\theta)$. We conclude by observing that, due to the continuity of prox_ψ (Lemma 2.2(iii)), this inequality holds for every ξ and η in $[-\rho, \rho]$. □

3. Proximal thresholding. The standard soft thresholder of (1.2), which was extended to closed intervals in (1.11), was seen in Example 2.6(ii) to be a proximity operator. As such, it possesses attractive properties (see Lemma 2.2(i) and (iii)) that prove extremely useful in the convergence analysis of iterative methods [13]. This remark motivates the following definition.

DEFINITION 3.1. *Let $R\colon \mathcal{H} \to \mathcal{H}$, and let Ω be a nonempty closed convex subset of $\mathcal{H}$. Then R is a* proximal thresholder *on Ω if there exists a function $f \in \Gamma_0(\mathcal{H})$ such that*

$$R = \operatorname{prox}_f \quad \textit{and} \quad (\forall x \in \mathcal{H}) \;\; Rx = 0 \;\Leftrightarrow\; x \in \Omega. \tag{3.1}$$

The next proposition provides characterizations of proximal thresholders.

PROPOSITION 3.2. *Let $f \in \Gamma_0(\mathcal{H})$, and let Ω be a nonempty closed convex subset of $\mathcal{H}$. Then the following are equivalent:*

(i) prox_f *is a proximal thresholder on Ω.*
(ii) $\partial f(0) = \Omega$.
(iii) $(\forall x \in \mathcal{H}) \; \big[\operatorname{prox}_{f^*} x = x \Leftrightarrow x \in \Omega \big]$.
(iv) $\operatorname{Argmin} f^* = \Omega$.

In particular, (i)–(iv) *hold when*

(v) $f = g + \sigma_\Omega$, *where $g \in \Gamma_0(\mathcal{H})$ is Gâteaux differentiable at 0 and $\nabla g(0) = 0$.*

Proof. (i)⇔(ii): Fix $x \in \mathcal{H}$. Then it follows from (2.4) that $\big[\operatorname{prox}_f x = 0 \Leftrightarrow x \in \Omega\big] \Leftrightarrow \big[x \in \partial f(0) \Leftrightarrow x \in \Omega\big] \Leftrightarrow \partial f(0) = \Omega$. (i)⇔(iii): Fix $x \in \mathcal{H}$. Then it follows from Lemma 2.2(ii) that $\big[\operatorname{prox}_f x = 0 \Leftrightarrow x \in \Omega\big] \Leftrightarrow \big[x - \operatorname{prox}_{f^*} x = 0 \Leftrightarrow x \in \Omega\big]$. (iii)⇔(iv): Since $f \in \Gamma_0(\mathcal{H})$, $f^* \in \Gamma_0(\mathcal{H})$, and we can apply Lemma 2.2(i) to f^*. (v)⇒(ii): Since (v) implies that $0 \in \operatorname{core} \operatorname{dom} g$, we have $0 \in (\operatorname{core} \operatorname{dom} g) \cap \operatorname{dom} \sigma_\Omega$, and it follows from [41, Theorem 2.8.3] that

$$\partial f(0) = \partial(g + \sigma_\Omega)(0) = \partial g(0) + \partial \sigma_\Omega(0) = \partial g(0) + \Omega, \tag{3.2}$$

where the last equality results from the observation that, for every $u \in \mathcal{H}$, Fenchel's identity yields $u \in \partial\sigma_\Omega(0) \Leftrightarrow 0 = \langle 0 \mid u\rangle = \sigma_\Omega(0) + \sigma_\Omega^*(u) \Leftrightarrow 0 = \sigma_\Omega^*(u) = \iota_\Omega(u) \Leftrightarrow u \in \Omega$. However, since $\partial g(0) = \{\nabla g(0)\} = \{0\}$, we obtain $\partial f(0) = \Omega$, and (ii) is therefore satisfied. □

The following theorem is a significant refinement of a result of Proposition 3.2 in the case when $\mathcal{H} = \mathbb{R}$ that characterizes all the functions $\phi \in \Gamma_0(\mathbb{R})$ for which prox_ϕ is a proximal thresholder.

THEOREM 3.3. *Let $\phi \in \Gamma_0(\mathbb{R})$ and let $\Omega \subset \mathbb{R}$ be a nonempty closed interval. Then the following are equivalent:*

(i) prox_ϕ *is a proximal thresholder on* Ω.

(ii) $\phi = \psi + \sigma_\Omega$, *where* $\psi \in \Gamma_0(\mathbb{R})$ *is differentiable at* 0 *and* $\psi'(0) = 0$.

Proof. In view of Proposition 3.2, it is enough to show that $\partial\phi(0) = \Omega \Rightarrow$ (ii). So let us assume that $\partial\phi(0) = \Omega$, and set $\underline{\omega} = \inf \Omega$ and $\overline{\omega} = \sup \Omega$. Since $\partial\phi(0) \neq \varnothing$, we deduce from (2.3) that $0 \in \operatorname{dom}\phi$ and that

$$(\forall \xi \in \mathbb{R}) \quad \sigma_\Omega(\xi) = \sup_{\nu \in \Omega} (\xi - 0)\nu \leq \phi(\xi) - \phi(0). \tag{3.3}$$

Consequently,

$$\operatorname{dom}\phi \subset \operatorname{dom}\sigma_\Omega. \tag{3.4}$$

Thus, in the case when $\Omega = \mathbb{R}$, Example 2.1(i) yields $\operatorname{dom}\phi = \operatorname{dom}\sigma_\Omega = \{0\}$ and we obtain $\phi = \phi(0) + \iota_{\{0\}} = \phi(0) + \sigma_\Omega$, hence (ii) with $\psi \equiv \phi(0)$. We henceforth assume that $\Omega \neq \mathbb{R}$ and set

$$(\forall \xi \in \mathbb{R}) \quad \varphi(\xi) = \begin{cases} \phi(\xi) - \phi(0) - \overline{\omega}\,\xi & \text{if } \xi > 0 \text{ and } \overline{\omega} < +\infty, \\ \phi(\xi) - \phi(0) - \underline{\omega}\,\xi & \text{if } \xi < 0 \text{ and } \underline{\omega} > -\infty, \\ 0 & \text{otherwise.} \end{cases} \tag{3.5}$$

Then Example 2.1(i) and (3.3) yield

$$\varphi \geq 0 = \varphi(0), \tag{3.6}$$

which also shows that φ is proper. In addition, we derive from Example 2.1(i) and (3.5) the following three possible expressions for φ:

(a) If $\underline{\omega} > -\infty$ and $\overline{\omega} < +\infty$, then σ_Ω is a finite continuous function and

$$(\forall \xi \in \mathbb{R}) \quad \varphi(\xi) = \phi(\xi) - \phi(0) - \sigma_\Omega(\xi). \tag{3.7}$$

(b) If $\underline{\omega} = -\infty$ and $\overline{\omega} < +\infty$, then

$$(\forall \xi \in \mathbb{R}) \quad \varphi(\xi) = \begin{cases} \phi(\xi) - \phi(0) - \overline{\omega}\,\xi & \text{if } \xi > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.8}$$

(c) If $\underline{\omega} > -\infty$ and $\overline{\omega} = +\infty$, then

$$(\forall \xi \in \mathbb{R}) \quad \varphi(\xi) = \begin{cases} \phi(\xi) - \phi(0) - \underline{\omega}\,\xi & \text{if } \xi < 0 \\ 0 & \text{otherwise.} \end{cases} \tag{3.9}$$

Let us show that φ is lower semicontinuous. In case (a), this follows at once from the lower semicontinuity of ϕ and the continuity of σ_Ω. In cases (b) and (c), φ is clearly lower semicontinuous at every point $\xi \neq 0$ and, by (3.6), at 0 as well. Next, let us establish the convexity of φ. To this end, we set

$$(\forall \xi \in \mathbb{R}) \quad \overline{\varphi}(\xi) = \begin{cases} \phi(\xi) - \phi(0) - \overline{\omega}\,\xi & \text{if } \xi > 0 \text{ and } \overline{\omega} < +\infty, \\ 0 & \text{otherwise,} \end{cases} \tag{3.10}$$

and

$$(\forall \xi \in \mathbb{R}) \quad \underline{\varphi}(\xi) = \begin{cases} \phi(\xi) - \phi(0) - \underline{\omega}\,\xi & \text{if } \xi < 0 \text{ and } \underline{\omega} > -\infty, \\ 0 & \text{otherwise.} \end{cases} \tag{3.11}$$

By inspecting (3.5), (3.10), and (3.11) we learn that φ coincides with $\overline{\varphi}$ on $[0, +\infty[$ and with $\underline{\varphi}$ on $]-\infty, 0]$. Hence, (3.6) yields

$$\overline{\varphi} \geq 0 \quad \text{and} \quad \underline{\varphi} \geq 0, \tag{3.12}$$

and

$$\varphi = \max\{\underline{\varphi}, \overline{\varphi}\}. \tag{3.13}$$

Furthermore, since ϕ is convex, so are the functions $\xi \mapsto \phi(\xi) - \phi(0) - \overline{\omega}\,\xi$ and $\xi \mapsto \phi(\xi) - \phi(0) - \underline{\omega}\,\xi$, when $\overline{\omega} < +\infty$ and $\underline{\omega} > -\infty$, respectively. Therefore, it follows from (3.10), (3.11), and (3.12) that $\overline{\varphi}$ and $\underline{\varphi}$ are convex, and hence from (3.13) that φ is convex. We have thus shown that $\varphi \in \Gamma_0(\mathbb{R})$. We now claim that, for every $\xi \in \mathbb{R}$,

$$\phi(\xi) = \varphi(\xi) + \phi(0) + \sigma_\Omega(\xi). \tag{3.14}$$

We can establish this identity with the help of Example 2.1(i). In case (a), (3.14) follows at once from (3.7) since σ_Ω is finite. In case (b), (3.14) follows from (3.8) when $\xi \geq 0$, and from (3.3) when $\xi < 0$ since, in this case, $\sigma_\Omega(\xi) = +\infty$. Likewise, in case (c), (3.14) follows from (3.9) when $\xi \leq 0$, and from (3.3) when $\xi > 0$ since, in this case, $\sigma_\Omega(\xi) = +\infty$. Next, let us show that

$$0 \in \operatorname{int}(\operatorname{dom}\phi - \operatorname{dom}\sigma_\Omega). \tag{3.15}$$

In case (a), we have $\Omega = [\underline{\omega}, \overline{\omega}]$. Therefore, $\operatorname{dom}\sigma_\Omega = \mathbb{R}$ and (3.15) trivially holds. In case (b), we have $\Omega =]-\infty, \overline{\omega}]$ and, therefore, $\operatorname{dom}\sigma_\Omega = [0, +\infty[$. This implies, via (3.4), that $\operatorname{dom}\phi \subset [0, +\infty[$. Therefore, there exists $\nu \in \operatorname{dom}\phi \cap]0, +\infty[$ since otherwise we would have $\operatorname{dom}\phi = \{0\}$, which, in view of (2.3), would contradict the current working assumption that $\partial\phi(0) = \Omega \neq \mathbb{R}$. By convexity of ϕ, it follows that $[0, \nu] \subset \operatorname{dom}\phi$ and, therefore, that $]-\infty, \nu] \subset \operatorname{dom}\phi - \operatorname{dom}\sigma_\Omega$. We thus obtain (3.15) in case (b); case (c) can be handled analogously. We can now appeal to [32, Theorem 23.8] to derive from (3.14), (3.15), and Example 2.1(ii) that

$$\Omega = \partial\phi(0) = \partial\varphi(0) + \partial\sigma_\Omega(0) = \partial\varphi(0) + \Omega. \tag{3.16}$$

Now fix $\nu \in \partial\varphi(0)$. Then (3.16) yields $\nu + \Omega \subset \Omega$. There are three possible cases to study.

- In case (a), $\nu+\Omega\subset\Omega\Leftrightarrow[\nu+\underline{\omega},\nu+\overline{\omega}]\subset[\underline{\omega},\overline{\omega}]\Rightarrow\nu=0$.
- In case (b), $\nu+\Omega\subset\Omega\Leftrightarrow\,]-\infty,\nu+\overline{\omega}]\subset\,]-\infty,\overline{\omega}]\Rightarrow\nu\leq 0$. On the other hand, it follows from (2.3) and (3.8) that $(\forall\xi\in\,]-\infty,0[)\ \xi\nu\leq\varphi(\xi)=0$, hence $\nu\geq 0$. Altogether, $\nu=0$.
- In case (c), $\nu+\Omega\subset\Omega\Leftrightarrow[\nu+\underline{\omega},+\infty[\,\subset[\underline{\omega},+\infty[\,\Rightarrow\nu\geq 0$. Since (2.3) and (3.9) imply that $(\forall\xi\in\,]0,+\infty[)\ \xi\nu\leq\varphi(\xi)=0$, we obtain $\nu\leq 0$ and conclude that $\nu=0$.

We have thus shown in all cases that $\nu=0$ and, therefore, that $\partial\varphi(0)=\{0\}$. In turn, upon invoking [32, Theorem 25.1], we conclude that φ is differentiable at 0 and that $\varphi'(0)=0$. Altogether, we obtain (ii) by setting $\psi=\varphi+\phi(0)$. □

Remark 3.4. A standard requirement for thresholders on $\mathbb{R}$ is that they be increasing functions [1, 33, 34, 39]. On the other hand, nonexpansivity is a key property to establish the convergence of iterative methods [13] and, in particular, in Proposition 1.1 [18] and Proposition 1.2 [16]. As seen in Proposition 2.4 and Definition 3.1, the increasing and nonexpansive functions $\varrho\colon\mathbb{R}\to\mathbb{R}$ that vanish only on a closed interval $\Omega\subset\mathbb{R}$ coincide with the proximal thresholders on Ω. Hence, appealing to Theorem 3.3 and Lemma 2.3, we conclude that the operators that perform a componentwise increasing and nonexpansive thresholding on $(\Omega_k)_{k\in\mathbb{K}}$ of those coefficients of the decomposition in $(e_k)_{k\in\mathbb{N}}$ indexed by $\mathbb{K}$ are precisely the operators of the form prox_Ψ, where Ψ is as in (1.9).

Example 3.5. Let $\omega\in\,]0,+\infty[$ and set

$$\phi\colon\mathbb{R}\to\,]-\infty,+\infty]:\xi\mapsto\begin{cases}\ln(\omega)-\ln(\omega-|\xi|) & \text{if } |\xi|<\omega,\\ +\infty & \text{otherwise.}\end{cases}\tag{3.17}$$

The proximity operator associated with this function arises in certain Bayesian formulations involving the triangular probability density function with support $[-\omega,\omega]$ [8]. Let us set

$$\psi\colon\mathbb{R}\to\,]-\infty,+\infty]:\xi\mapsto\begin{cases}\ln(\omega)-\ln(\omega-|\xi|)-|\xi|/\omega & \text{if } |\xi|<\omega,\\ +\infty & \text{otherwise}\end{cases}\tag{3.18}$$

and $\Omega=[-1/\omega,1/\omega]$. Then $\psi\in\Gamma_0(\mathbb{R})$ is differentiable at 0, $\psi'(0)=0$, and $\phi=\psi+\sigma_\Omega$. Therefore, Theorem 3.3 asserts that prox_ϕ is a proximal thresholder on $[-1/\omega,1/\omega]$. Actually (see Figure 3.1), for every $\xi\in\mathbb{R}$, we have [8, Example 4.12]

$$\operatorname{prox}_\phi\xi=\begin{cases}\operatorname{sign}(\xi)\,\dfrac{|\xi|+\omega-\sqrt{\big||\xi|-\omega\big|^2+4}}{2} & \text{if } |\xi|>1/\omega,\\ 0 & \text{otherwise.}\end{cases}\tag{3.19}$$

Next, we provide a convenient decomposition rule for implementing proximal thresholders.

PROPOSITION 3.6. *Let $\phi=\psi+\sigma_\Omega$, where $\psi\in\Gamma_0(\mathbb{R})$ and $\Omega\subset\mathbb{R}$ is a nonempty closed interval. Suppose that ψ is differentiable at 0 with $\psi'(0)=0$. Then* $\operatorname{prox}_\phi=\operatorname{prox}_\psi\circ\operatorname{soft}_\Omega$.

Proof. Fix ξ and π in $\mathbb{R}$. We have $0\in\operatorname{dom}\sigma_\Omega$ and, since ψ is differentiable at 0, $0\in\operatorname{int}\operatorname{dom}\psi$. It therefore follows from (2.4) and [32, Theorem 23.8] that

$$\begin{aligned}\pi=\operatorname{prox}_\phi\xi&\Leftrightarrow\xi-\pi\in\partial\phi(\pi)=\partial\psi(\pi)+\partial\sigma_\Omega(\pi)\\ &\Leftrightarrow(\exists\,\nu\in\partial\psi(\pi))\quad\xi-(\pi+\nu)\in\partial\sigma_\Omega(\pi).\end{aligned}\tag{3.20}$$

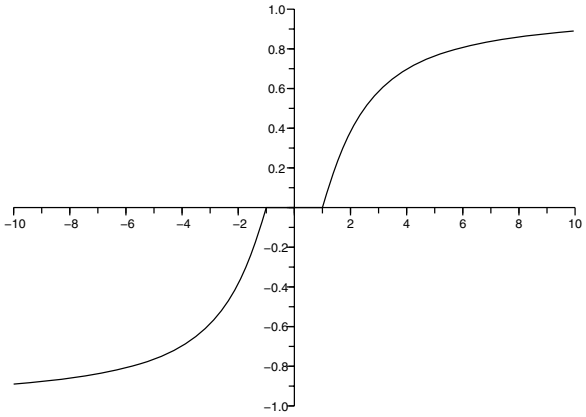


FIG. 3.1. *Graph of* prox_{ϕ}*, where* ϕ *is as in* (3.17) *with* $\omega = 1$.

Let us observe that, if $\nu \in \partial\psi(\pi)$, then, since $0 \in \operatorname{Argmin}\psi$, (2.3) implies that $(0-\pi)\nu + \psi(\pi) \leq \psi(0) \leq \psi(\pi) < +\infty$ and, in turn, that $\pi\nu \geq 0$. This shows that, if $\nu \in \partial\psi(\pi)$ and $\pi \neq 0$, then either $\pi > 0$ and $\nu \geq 0$, or $\pi < 0$ and $\nu \leq 0$; in turn, Lemma 2.1(ii) yields $\partial\sigma_{\Omega}(\pi) = \partial\sigma_{\Omega}(\pi+\nu)$. Consequently, if $\pi \neq 0$, we derive from (3.20) and Example 2.6(ii) that

$$\begin{aligned}
\pi = \mathrm{prox}_{\phi}\,\xi \Rightarrow (\exists\,\nu \in \partial\psi(\pi)) \quad & \xi - (\pi+\nu) \in \partial\sigma_{\Omega}(\pi+\nu) \\
\Leftrightarrow (\exists\,\nu \in \partial\psi(\pi)) \quad & \pi + \nu = \mathrm{prox}_{\sigma_{\Omega}}\,\xi = \mathrm{soft}_{\Omega}\,\xi \\
\Leftrightarrow\ & \mathrm{soft}_{\Omega}\,\xi - \pi \in \partial\psi(\pi) \\
\Leftrightarrow\ & \pi = \mathrm{prox}_{\psi}\big(\mathrm{soft}_{\Omega}\,\xi\big).
\end{aligned} \tag{3.21}$$

On the other hand, if $\pi = 0$, since $\partial\psi(0) = \{\psi'(0)\} = \{0\}$, then we derive from (3.20), Example 2.1(ii), (1.11), and Lemma 2.2(i) that

$$\pi = \mathrm{prox}_{\phi}\,\xi \Rightarrow \xi \in \partial\sigma_{\Omega}(0) = \Omega \Rightarrow \mathrm{soft}_{\Omega}\,\xi = 0 \Rightarrow \mathrm{prox}_{\psi}\big(\mathrm{soft}_{\Omega}\,\xi\big) = 0 = \pi. \tag{3.22}$$

The proof is now complete. □

In view of Proposition 3.6 and (1.11), the computation of the proximal thresholder $\mathrm{prox}_{\psi+\sigma_{\Omega}}$ reduces to that of prox_{ψ}. By duality, we obtain a decomposition formula for those proximal operators that coincide with the identity on a closed interval Ω.

PROPOSITION 3.7. *Let* $\phi = \psi \,\square\, \iota_{\Omega}$*, where* $\psi \in \Gamma_0(\mathbb{R})$ *and* $\Omega \subset \mathbb{R}$ *is a nonempty closed interval. Suppose that* ψ^* *is differentiable at* 0 *with* $\psi^{*\prime}(0) = 0$*. Then the following hold.*

(i) $\mathrm{prox}_{\phi} = P_{\Omega} + \mathrm{prox}_{\psi} \circ \mathrm{soft}_{\Omega}$.
(ii) $(\forall \xi \in \mathbb{R})\ \mathrm{prox}_{\phi}\,\xi = \xi \Leftrightarrow \xi \in \Omega$.

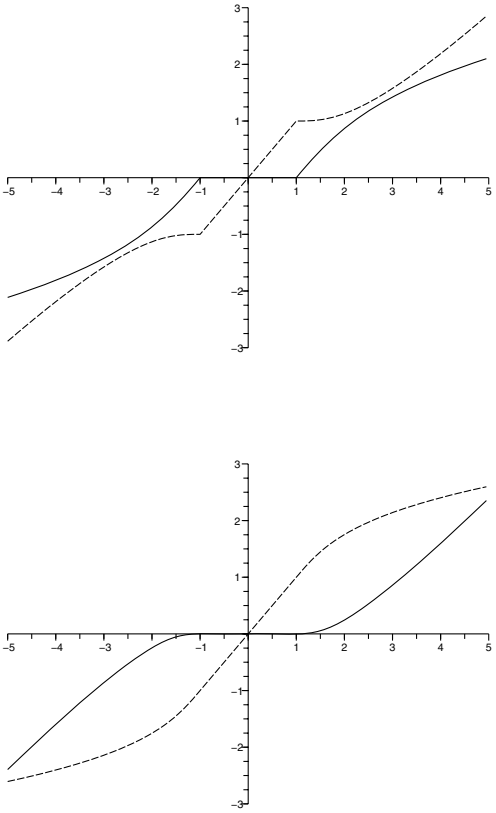


FIG. 3.2. *Graphs of the proximal thresholder* prox_ϕ *(solid line) and its dual* $\operatorname{prox}_{\phi^*}$ *(dashed line), where* $\phi = \tau|\cdot|^p + |\cdot|$. *Top:* $\tau = 0.05$ *and* $p = 4$. *Bottom:* $\tau = 0.9$ *and* $p = 4/3$. *Explicit expressions for these thresholders are provided by Example* 2.7(ii) *and* (vi), *Proposition* 3.6, *and Lemma* 2.2(ii).

Proof. It follows from [32, Theorem 16.4] that

$$\phi^* = \psi^* + \iota_\Omega^* = \psi^* + \sigma_\Omega. \tag{3.23}$$

Note also that, since $\psi \in \Gamma_0(\mathbb{R})$, we have $\psi^* \in \Gamma_0(\mathbb{R})$ [32, Theorem 12.2]. (i) Fix $\xi \in \mathbb{R}$. Then, by Lemma 2.2(ii), (3.23), Proposition 3.6, and Example 2.6,

$$\begin{aligned}
\operatorname{prox}_\phi \xi &= \xi - \operatorname{prox}_{\phi^*} \xi \\
&= \xi - \operatorname{prox}_{\psi^*+\sigma_\Omega} \xi \\
&= \xi - \operatorname{prox}_{\psi^*} \big(\operatorname{prox}_{\sigma_\Omega} \xi\big) \\
&= \xi - \operatorname{prox}_{\sigma_\Omega} \xi + \operatorname{prox}_{\psi} \big(\operatorname{prox}_{\sigma_\Omega} \xi\big) \\
&= \operatorname{prox}_{\sigma_\Omega^*} \xi + \operatorname{prox}_{\psi} \big(\operatorname{prox}_{\sigma_\Omega} \xi\big) \\
&= \operatorname{prox}_{\iota_\Omega} \xi + \operatorname{prox}_{\psi} \big(\operatorname{prox}_{\sigma_\Omega} \xi\big) \qquad (3.24) \\
&= P_\Omega \xi + \operatorname{prox}_{\psi} \big(\operatorname{soft}_\Omega \xi\big). \qquad (3.25)
\end{aligned}$$

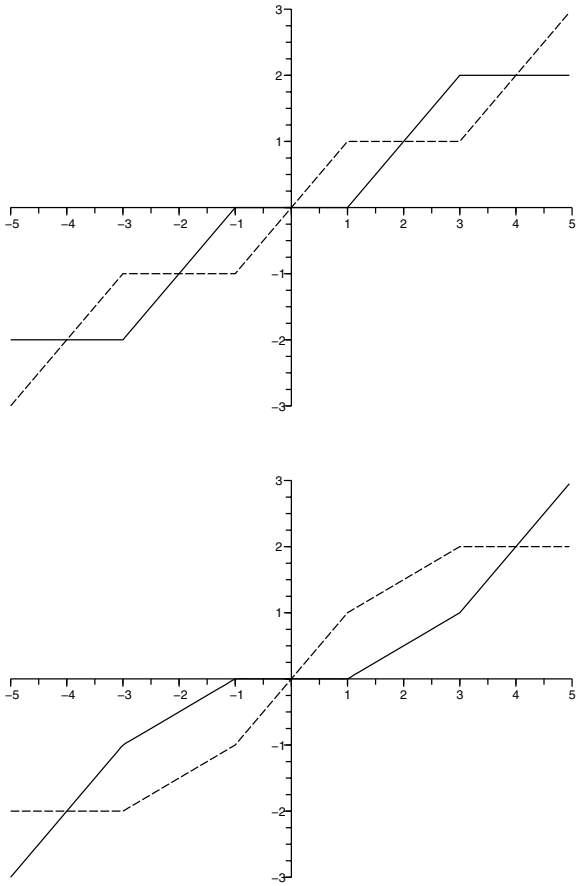


FIG. 3.3. *Graphs of the proximal thresholder* prox_ϕ *(solid line) and its dual* $\operatorname{prox}_{\phi^*}$ *(dashed line), where* $\phi = \psi + |\cdot|$. *Top:* $\psi = \iota_{[-2,2]}$. *Bottom:* $\psi\colon \xi \mapsto \xi^2/2$, *if* $|\xi| \leq 1; |\xi| - 1/2$, *if* $|\xi| > 1$ *is the Huber function* [27]. *The closed-form expressions of these thresholders are obtained via* [8, Example 4.5], *Proposition* 3.6, *and Lemma* 2.2(ii).

(ii): It follows from (3.23) and Theorem 3.3 that $\operatorname{prox}_{\phi^*}$ is a proximal thresholder on Ω. Hence, we derive from (3.24) and (3.1) that $(\forall \xi \in \mathbb{R})\ \operatorname{prox}_\phi \xi = \xi \Leftrightarrow \operatorname{prox}_{\phi^*} \xi = 0 \Leftrightarrow \xi \in \Omega$. □

Examples of proximal thresholders (see Proposition 3.6) and their duals (see Proposition 3.7) are provided in Figures 3.2 and 3.3 (see also Figure 2.1) in the case when $\Omega = [-1, 1]$.

4. Iterative proximal thresholding. Let us start with some basic properties of Problem 1.3.

PROPOSITION 4.1. *Problem* 1.3 *possesses at least one solution.*

Proof. Let Ψ be as in (1.9). We infer from the assumptions of Problem 1.3 and Lemma 2.3 that $\Psi \in \Gamma_0(\mathcal{H})$ and, in turn, that $\Phi + \Psi \in \Gamma_0(\mathcal{H})$. Hence, it suffices to show that $\Phi + \Psi$ is coercive [41, Theorem 2.5.1(ii)], i.e., since $\inf \Phi(\mathcal{H}) > -\infty$ by

assumption (i) in Problem 1.3, that Ψ is coercive. For this purpose, let $\mathsf{x} = (\xi_k)_{k\in\mathbb{N}}$ denote a generic element in $\ell^2(\mathbb{N})$, and let

$$\Upsilon\colon \ell^2(\mathbb{N}) \to \left]-\infty,+\infty\right] : \mathsf{x} \mapsto \sum_{k\in\mathbb{N}} \psi_k(\xi_k) + \sum_{k\in\mathbb{K}} \sigma_{\Omega_k}(\xi_k). \tag{4.1}$$

Then, by Parseval's identity, it is enough to show that Υ is coercive. To this end, set $\mathsf{x}_{\mathbb{K}} = (\xi_k)_{k\in\mathbb{K}}$ and $\mathsf{x}_{\mathbb{L}} = (\xi_k)_{k\in\mathbb{L}}$, and denote by $\|\cdot\|_{\mathbb{K}}$ and $\|\cdot\|_{\mathbb{L}}$ the standard norms on $\ell^2(\mathbb{K})$ and $\ell^2(\mathbb{L})$, respectively. Using (4.1), assumptions (ii) and (vi) in Problem 1.3, and Example 2.1(i), we obtain

$$\begin{aligned}(\forall\, \mathsf{x} \in \ell^2(\mathbb{N}))\quad \Upsilon(\mathsf{x}) &\geq \sum_{k\in\mathbb{K}} \sigma_{\Omega_k}(\xi_k) + \sum_{k\in\mathbb{L}} \psi_k(\xi_k)\\ &\geq \omega \sum_{k\in\mathbb{K}} |\xi_k| + \Upsilon_{\mathbb{L}}(\mathsf{x}_{\mathbb{L}})\\ &\geq \omega\|\mathsf{x}_{\mathbb{K}}\|_{\mathbb{K}} + \Upsilon_{\mathbb{L}}(\mathsf{x}_{\mathbb{L}}),\end{aligned} \tag{4.2}$$

where $\Upsilon_{\mathbb{L}}$ is defined in Problem 1.3(v). Now suppose that $\|\mathsf{x}\| = \sqrt{\|\mathsf{x}_{\mathbb{K}}\|_{\mathbb{K}}^2 + \|\mathsf{x}_{\mathbb{L}}\|_{\mathbb{L}}^2} \to +\infty$. Then (4.2) and assumption (v) in Problem 1.3 yield $\Upsilon(\mathsf{x}) \to +\infty$, as desired. ∎

PROPOSITION 4.2. *Let Ψ be as in (1.9), let $x \in \mathcal{H}$, and let $\gamma \in \left]0,+\infty\right[$. Then x is a solution to Problem 1.3 if and only if $x = \operatorname{prox}_{\gamma\Psi}(x - \gamma\nabla\Phi(x))$.*

Proof. Since Problem 1.3 is equivalent to minimizing $\Phi + \Psi$, this identity is a standard characterization; see, for instance, [16, Proposition 3.1(iii)]. ∎

Our algorithm for solving Problem 1.3 will be the following.

ALGORITHM 4.3. *Fix $x_0 \in \mathcal{H}$ and set, for every $n \in \mathbb{N}$,*

$$\begin{aligned}x_{n+1} = x_n + \lambda_n \Bigg(&\sum_{k\in\mathbb{K}} \Big(\alpha_{n,k} + \operatorname{prox}_{\gamma_n\psi_k}\big(\operatorname{soft}_{\gamma_n\Omega_k}\langle x_n - \gamma_n(\nabla\Phi(x_n) + b_n) \mid e_k\rangle\big)\Big)e_k\\ &+ \sum_{k\in\mathbb{L}} \Big(\alpha_{n,k} + \operatorname{prox}_{\gamma_n\psi_k}\langle x_n - \gamma_n(\nabla\Phi(x_n) + b_n) \mid e_k\rangle\Big)e_k - x_n\Bigg),\end{aligned} \tag{4.3}$$

where

(i) *$(\gamma_n)_{n\in\mathbb{N}}$ is a sequence in $\left]0,+\infty\right[$ such that $\inf_{n\in\mathbb{N}}\gamma_n > 0$ and $\sup_{n\in\mathbb{N}}\gamma_n < 2\beta$,*

(ii) *$(\lambda_n)_{n\in\mathbb{N}}$ is a sequence in $\left]0,1\right]$ such that $\inf_{n\in\mathbb{N}}\lambda_n > 0$,*

(iii) *for every $n \in \mathbb{N}$, $(\alpha_{n,k})_{k\in\mathbb{N}}$ is a sequence in $\ell^2(\mathbb{N})$ such that*

$$\sum_{n\in\mathbb{N}} \sqrt{\sum_{k\in\mathbb{N}} |\alpha_{n,k}|^2} < +\infty,$$

(iv) *$(b_n)_{n\in\mathbb{N}}$ is a sequence in $\mathcal{H}$ such that $\sum_{n\in\mathbb{N}}\|b_n\| < +\infty$.*

Remark 4.4. Let us highlight some features of Algorithm 4.3.

- The set $\mathbb{K}$ contains the indices of those coefficients of the decomposition in $(e_k)_{k\in\mathbb{N}}$ that are thresholded.
- The terms $\alpha_{n,k}$ and b_n stand for some numerical tolerance in the implementation of $\operatorname{prox}_{\gamma_n\psi_k}$ and the computation of $\nabla\Phi(x_n)$, respectively.
- The parameters λ_n and γ_n provide added flexibility to the algorithm and can be used to improve its convergence profile.

- The operator $\mathrm{soft}_{\gamma_n \Omega_k}$ is given explicitly in (1.11).

Our main convergence result can now be stated.

THEOREM 4.5. *Every sequence generated by Algorithm* 4.3 *converges strongly to a solution to Problem* 1.3.

Proof. Hereafter, $(x_n)_{n\in\mathbb{N}}$ is a sequence generated by Algorithm 4.3, and we define

$$(\forall k \in \mathbb{N}) \quad \phi_k = \begin{cases} \psi_k + \sigma_{\Omega_k} & \text{if } k \in \mathbb{K}, \\ \psi_k & \text{if } k \in \mathbb{L}. \end{cases} \tag{4.4}$$

It follows from the assumptions on $(\psi_k)_{k\in\mathbb{N}}$ in Problem 1.3 that $(\forall k \in \mathbb{N})\ \psi_k'(0) = 0$. Therefore, for every n in $\mathbb{N}$, Theorem 3.3 implies that

$$\text{for every } k \text{ in } \mathbb{K},\ \mathrm{prox}_{\gamma_n \phi_k} \text{ is a proximal thresholder on } \gamma_n \Omega_k, \tag{4.5}$$

while Proposition 3.6 supplies

$$(\forall k \in \mathbb{K}) \quad \mathrm{prox}_{\gamma_n \phi_k} = \mathrm{prox}_{\gamma_n \psi_k + \gamma_n \sigma_{\Omega_k}} = \mathrm{prox}_{\gamma_n \psi_k + \sigma_{(\gamma_n \Omega_k)}} = \mathrm{prox}_{\gamma_n \psi_k} \circ \mathrm{soft}_{\gamma_n \Omega_k}\,. \tag{4.6}$$

Thus, (4.3) can be rewritten as

$$x_{n+1} = x_n + \lambda_n \Bigg(\sum_{k\in\mathbb{N}} \big(\alpha_{n,k} + \mathrm{prox}_{\gamma_n \phi_k} \langle x_n - \gamma_n (\nabla \Phi(x_n) + b_n) \mid e_k \rangle\big) e_k - x_n \Bigg). \tag{4.7}$$

Now let Ψ be as in (1.9), i.e., $\Psi = \sum_{k\in\mathbb{N}} \phi_k(\langle \cdot \mid e_k \rangle)$, and set $(\forall n \in \mathbb{N})\ a_n = \sum_{k\in\mathbb{N}} \alpha_{n,k} e_k$. Then it follows from (4.4) and Lemma 2.3 that $\Psi \in \Gamma_0(\mathcal{H})$ and that (4.7) can be rewritten as

$$x_{n+1} = x_n + \lambda_n \Big(\mathrm{prox}_{\gamma_n \Psi} \big(x_n - \gamma_n (\nabla \Phi(x_n) + b_n)\big) + a_n - x_n \Big). \tag{4.8}$$

Consequently, since Proposition 4.1 asserts that $\Phi + \Psi$ possesses a minimizer, we derive from assumptions (i)–(iv) in Algorithm 4.3 and [16, Theorem 3.4] that

$$(x_n)_{n\in\mathbb{N}} \text{ converges weakly to a solution } x \text{ to Problem 1.3} \tag{4.9}$$

and that

$$\sum_{n\in\mathbb{N}} \|x_n - \mathrm{prox}_{\gamma_n \Psi} \big(x_n - \gamma_n \nabla \Phi(x_n)\big)\|^2 < +\infty \quad \text{and} \quad \sum_{n\in\mathbb{N}} \|\nabla \Phi(x_n) - \nabla \Phi(x)\|^2 < +\infty. \tag{4.10}$$

Hence, it follows from Lemma 2.2(iii) and assumption (i) in Algorithm 4.3 that

$$\begin{aligned}
(4.11)\quad & \frac{1}{2}\sum_{n\in\mathbb{N}} \|x_n - \operatorname{prox}_{\gamma_n\Psi}\big(x_n - \gamma_n\nabla\Phi(x)\big)\|^2 \\
&\leq \sum_{n\in\mathbb{N}} \|x_n - \operatorname{prox}_{\gamma_n\Psi}\big(x_n - \gamma_n\nabla\Phi(x_n)\big)\|^2 \\
&\quad + \sum_{n\in\mathbb{N}} \|\operatorname{prox}_{\gamma_n\Psi}\big(x_n - \gamma_n\nabla\Phi(x_n)\big) - \operatorname{prox}_{\gamma_n\Psi}\big(x_n - \gamma_n\nabla\Phi(x)\big)\|^2 \\
&\leq \sum_{n\in\mathbb{N}} \|x_n - \operatorname{prox}_{\gamma_n\Psi}\big(x_n - \gamma_n\nabla\Phi(x_n)\big)\|^2 + \sum_{n\in\mathbb{N}} \gamma_n^2\|\nabla\Phi(x_n) - \nabla\Phi(x)\|^2 \\
&\leq \sum_{n\in\mathbb{N}} \|x_n - \operatorname{prox}_{\gamma_n\Psi}\big(x_n - \gamma_n\nabla\Phi(x_n)\big)\|^2 + 4\beta^2\sum_{n\in\mathbb{N}} \|\nabla\Phi(x_n) - \nabla\Phi(x)\|^2 \\
&< +\infty.
\end{aligned}$$

Now define

$$(4.12)\qquad (\forall n\in\mathbb{N})\quad v_n = x_n - x \quad\text{and}\quad h_n = x - \gamma_n\nabla\Phi(x).$$

On the one hand, we derive from (4.9) that

$$(4.13)\qquad (v_n)_{n\in\mathbb{N}} \text{ converges weakly to } 0$$

and, on the other hand, we derive from (4.11) and Proposition 4.2 that

$$\begin{aligned}
\sum_{n\in\mathbb{N}} \|v_n - \operatorname{prox}_{\gamma_n\Psi}(v_n + h_n) + \operatorname{prox}_{\gamma_n\Psi} h_n\|^2 &= \sum_{n\in\mathbb{N}} \|x_n - \operatorname{prox}_{\gamma_n\Psi}\big(x_n - \gamma_n\nabla\Phi(x)\big)\|^2 \\
&< +\infty. \qquad (4.14)
\end{aligned}$$

By Parseval's identity, to establish that $\|v_n\| = \|x_n - x\| \to 0$, we must show that

$$(4.15)\qquad \sum_{k\in\mathbb{K}} |\nu_{n,k}|^2 \to 0 \quad\text{and}\quad \sum_{k\in\mathbb{L}} |\nu_{n,k}|^2 \to 0,$$

where $(\forall n\in\mathbb{N})(\forall k\in\mathbb{N})\ \nu_{n,k} = \langle v_n \mid e_k\rangle$. To this end, let us set, for every $n\in\mathbb{N}$ and $k\in\mathbb{N}$, $\eta_{n,k} = \langle h_n \mid e_k\rangle$ and observe that (4.14), Parseval's identity, and Lemma 2.3 imply that

$$(4.16)\qquad \sum_{k\in\mathbb{N}} |\nu_{n,k} - \operatorname{prox}_{\gamma_n\phi_k}(\nu_{n,k} + \eta_{n,k}) + \operatorname{prox}_{\gamma_n\phi_k}\eta_{n,k}|^2 \to 0.$$

In addition, let us set $r = 2\beta\nabla\Phi(x)$ and, for every $k\in\mathbb{N}$, $\xi_k = \langle x \mid e_k\rangle$ and $\rho_k = \langle r \mid e_k\rangle$. Then we derive from (4.12) and assumption (i) in Algorithm 4.3 that

$$(4.17)\qquad (\forall n\in\mathbb{N})(\forall k\in\mathbb{N})\quad |\eta_{n,k}|^2/2 \leq |\xi_k|^2 + \gamma_n^2\,|\langle\nabla\Phi(x) \mid e_k\rangle|^2 \leq |\xi_k|^2 + |\rho_k|^2.$$

To establish (4.15), let us first show that $\sum_{k\in\mathbb{K}} |\nu_{n,k}|^2 \to 0$. For this purpose, set $\delta = \gamma\omega$, where $\gamma = \inf_{n\in\mathbb{N}}\gamma_n$ and where ω is supplied by assumption (vi) in

Problem 1.3. Then it follows from assumption (i) in Algorithm 4.3 that $\delta > 0$ and that

$$[-\delta, \delta] \subset \bigcap_{n\in\mathbb{N}} \bigcap_{k\in\mathbb{K}} \gamma_n \Omega_k. \tag{4.18}$$

On the other hand, (4.17) yields

$$\sum_{k\in\mathbb{K}} \sup_{n\in\mathbb{N}} |\eta_{n,k}|^2/2 \leq \sum_{k\in\mathbb{N}} \big(|\xi_k|^2 + |\rho_k|^2\big) = \|x\|^2 + \|r\|^2 < +\infty. \tag{4.19}$$

Hence, there exists a finite set $\mathbb{K}_1 \subset \mathbb{K}$ such that

$$(\forall n \in \mathbb{N}) \quad \sum_{k\in\mathbb{K}_2} |\eta_{n,k}|^2 \leq \delta^2/4, \quad \text{where} \quad \mathbb{K}_2 = \mathbb{K} \smallsetminus \mathbb{K}_1. \tag{4.20}$$

In view of (4.13), we have $\sum_{k\in\mathbb{K}_1} |\nu_{n,k}|^2 \to 0$. Let us now show that $\sum_{k\in\mathbb{K}_2} |\nu_{n,k}|^2 \to 0$. Note that (4.18) and (4.20) yield

$$(\forall n \in \mathbb{N})(\forall k \in \mathbb{K}_2) \quad \eta_{n,k} \in [-\delta/2, \delta/2] \subset \gamma_n \Omega_k. \tag{4.21}$$

Therefore, (4.5) implies that

$$(\forall n \in \mathbb{N})(\forall k \in \mathbb{K}_2) \quad \operatorname{prox}_{\gamma_n \phi_k} \eta_{n,k} = 0. \tag{4.22}$$

Let us define

$$(\forall n \in \mathbb{N}) \quad \mathbb{K}_{21,n} = \big\{k \in \mathbb{K}_2 \;\big|\; \nu_{n,k} + \eta_{n,k} \in \gamma_n \Omega_k\big\}. \tag{4.23}$$

Then, invoking (4.5) once again, we obtain

$$(\forall n \in \mathbb{N})(\forall k \in \mathbb{K}_{21,n}) \quad \operatorname{prox}_{\gamma_n \phi_k}(\nu_{n,k} + \eta_{n,k}) = 0 \tag{4.24}$$

which, combined with (4.22), yields

$$\begin{aligned}(\forall n \in \mathbb{N}) \quad \sum_{k\in\mathbb{K}_{21,n}} |\nu_{n,k}|^2 &= \sum_{k\in\mathbb{K}_{21,n}} |\nu_{n,k} - \operatorname{prox}_{\gamma_n \phi_k}(\nu_{n,k} + \eta_{n,k}) + \operatorname{prox}_{\gamma_n \phi_k} \eta_{n,k}|^2 \\ &\leq \sum_{k\in\mathbb{N}} |\nu_{n,k} - \operatorname{prox}_{\gamma_n \phi_k}(\nu_{n,k} + \eta_{n,k}) + \operatorname{prox}_{\gamma_n \phi_k} \eta_{n,k}|^2. \end{aligned} \tag{4.25}$$

Consequently, it results from (4.16) that $\sum_{k\in\mathbb{K}_{21,n}} |\nu_{n,k}|^2 \to 0$. Next, let us set

$$(\forall n \in \mathbb{N}) \quad \mathbb{K}_{22,n} = \mathbb{K}_2 \smallsetminus \mathbb{K}_{21,n} \tag{4.26}$$

and show that $\sum_{k\in\mathbb{K}_{22,n}} |\nu_{n,k}|^2 \to 0$. It follows from (4.26), (4.23), and (4.18) that

$$(\forall n \in \mathbb{N})(\forall k \in \mathbb{K}_{22,n}) \quad \nu_{n,k} + \eta_{n,k} \notin \gamma_n \Omega_k \supset [-\delta, \delta]. \tag{4.27}$$

Hence, appealing to (4.21), we obtain

$$(\forall n \in \mathbb{N})(\forall k \in \mathbb{K}_{22,n}) \quad |\nu_{n,k} + \eta_{n,k}| \geq \delta \geq |\eta_{n,k}| + \delta/2. \tag{4.28}$$

Now take $n \in \mathbb{N}$ and $k \in \mathbb{K}_{22,n}$. We derive from (4.22) and Lemma 2.2(ii) that

$$\begin{aligned}(4.29)\quad &|\nu_{n,k} - \operatorname{prox}_{\gamma_n\phi_k}(\nu_{n,k}+\eta_{n,k}) + \operatorname{prox}_{\gamma_n\phi_k}\eta_{n,k}| \\ &\qquad = |(\nu_{n,k}+\eta_{n,k}) - \operatorname{prox}_{\gamma_n\phi_k}(\nu_{n,k}+\eta_{n,k}) - \eta_{n,k}| \\ &\qquad = |\operatorname{prox}_{(\gamma_n\phi_k)^*}(\nu_{n,k}+\eta_{n,k}) - \eta_{n,k}|.\end{aligned}$$

However, it results from (4.18), (4.5), and Proposition 3.2 that $\operatorname{prox}_{(\gamma_n\phi_k)^*}(\pm\delta) = \pm\delta$. We consider two cases. First, if $\nu_{n,k}+\eta_{n,k} \geq 0$, then, since $\operatorname{prox}_{(\gamma_n\phi_k)^*}$ is increasing by Proposition 2.4, (4.28) yields $\nu_{n,k}+\eta_{n,k} \geq \delta$ and

$$(4.30)\qquad \operatorname{prox}_{(\gamma_n\phi_k)^*}(\nu_{n,k}+\eta_{n,k}) \geq \operatorname{prox}_{(\gamma_n\phi_k)^*}\delta = \delta \geq \eta_{n,k} + \delta/2.$$

Likewise, if $\nu_{n,k}+\eta_{n,k} \leq 0$, then (4.28) yields $\nu_{n,k}+\eta_{n,k} \leq -\delta$ and

$$(4.31)\qquad \operatorname{prox}_{(\gamma_n\phi_k)^*}(\nu_{n,k}+\eta_{n,k}) \leq \operatorname{prox}_{(\gamma_n\phi_k)^*}(-\delta) = -\delta \leq \eta_{n,k} - \delta/2.$$

Altogether, we derive from (4.30) and (4.31) that

$$(4.32)\qquad (\forall n \in \mathbb{N})(\forall k \in \mathbb{K}_{22,n}) \quad |\operatorname{prox}_{(\gamma_n\phi_k)^*}(\nu_{n,k}+\eta_{n,k}) - \eta_{n,k}| \geq \delta/2.$$

In turn, (4.29) yields

$$(4.33)\qquad (\forall n \in \mathbb{N}) \quad \sum_{k\in\mathbb{K}_{22,n}} |\nu_{n,k} - \operatorname{prox}_{\gamma_n\phi_k}(\nu_{n,k}+\eta_{n,k}) + \operatorname{prox}_{\gamma_n\phi_k}\eta_{n,k}|^2 \geq \operatorname{card}(\mathbb{K}_{22,n})\delta^2/4.$$

However, it follows from (4.16) that, for n sufficiently large,

$$(4.34)\qquad \sum_{k\in\mathbb{N}} |\nu_{n,k} - \operatorname{prox}_{\gamma_n\phi_k}(\nu_{n,k}+\eta_{n,k}) + \operatorname{prox}_{\gamma_n\phi_k}\eta_{n,k}|^2 \leq \delta^2/5.$$

Thus, for n sufficiently large, $\mathbb{K}_{22,n} = \varnothing$. We conclude from this first part of the proof that $\sum_{k\in\mathbb{K}} |\nu_{n,k}|^2 \to 0$.

In order to obtain (4.15), we must now show that $\sum_{k\in\mathbb{L}} |\nu_{n,k}|^2 \to 0$. We infer from (4.13) that $(v_n)_{n\in\mathbb{N}}$ is bounded; hence

$$(4.35)\qquad \sup_{n\in\mathbb{N}} \sum_{k\in\mathbb{L}} |\nu_{n,k}|^2 \leq \sup_{n\in\mathbb{N}} \|v_n\|^2 \leq \rho^2/4$$

for some $\rho \in \left]0, +\infty\right[$. Now define

$$(4.36)\qquad \mathbb{L}_1 = \big\{k \in \mathbb{L} \;\big|\; (\exists\, n \in \mathbb{N}) \;\; |\eta_{n,k}| \geq \rho/2\big\}.$$

Then we derive from (4.17) that

$$(4.37)\qquad (\forall k \in \mathbb{L}_1)(\exists\, n \in \mathbb{N}) \quad |\xi_k|^2 + |\rho_k|^2 \geq |\eta_{n,k}|^2/2 \geq \rho^2/8.$$

Consequently, we have

$$(4.38)\qquad +\infty > \|x\|^2 + \|r\|^2 \geq \sum_{k\in\mathbb{L}_1} \big(|\xi_k|^2 + |\rho_k|^2\big) \geq (\operatorname{card}\mathbb{L}_1)\rho^2/8$$

and therefore $\operatorname{card}(\mathbb{L}_1) < +\infty$. In turn, it results from (4.13) that $\sum_{k\in\mathbb{L}_1} |\nu_{n,k}|^2 \to 0$. Hence, to obtain $\sum_{k\in\mathbb{L}} |\nu_{n,k}|^2 \to 0$, it remains to show that $\sum_{k\in\mathbb{L}_2} |\nu_{n,k}|^2 \to 0$, where $\mathbb{L}_2 = \mathbb{L} \smallsetminus \mathbb{L}_1$. In view of (4.36) and (4.35), we have

$$(\forall n \in \mathbb{N})(\forall k \in \mathbb{L}_2) \quad |\eta_{n,k}| < \rho/2 \quad \text{and} \quad |\nu_{n,k} + \eta_{n,k}| \leq |\nu_{n,k}| + |\eta_{n,k}| < \rho. \tag{4.39}$$

On the other hand, assumption (iv) in Problem 1.3 asserts that there exists $\theta \in]0, +\infty[$ such that

$$\inf_{n\in\mathbb{N}} \inf_{k\in\mathbb{L}_2} \inf_{0<|\xi|\leq\rho} (\gamma_n\psi_k)''(\xi) \geq \gamma \inf_{k\in\mathbb{L}_2} \inf_{0<|\xi|\leq\rho} \psi_k''(\xi) \geq \gamma\theta. \tag{4.40}$$

It therefore follows from assumptions (ii) and (iii) in Problem 1.3, Proposition 2.9, and (4.4) that

$$\begin{aligned}
(\forall n \in \mathbb{N})(\forall k \in \mathbb{L}_2) \quad |\nu_{n,k}| &\leq |\nu_{n,k} - \operatorname{prox}_{\gamma_n\psi_k}(\nu_{n,k} + \eta_{n,k}) + \operatorname{prox}_{\gamma_n\psi_k} \eta_{n,k}| \\
&\quad + |\operatorname{prox}_{\gamma_n\psi_k}(\nu_{n,k} + \eta_{n,k}) - \operatorname{prox}_{\gamma_n\psi_k} \eta_{n,k}| \\
&\leq |\nu_{n,k} - \operatorname{prox}_{\gamma_n\psi_k}(\nu_{n,k} + \eta_{n,k}) + \operatorname{prox}_{\gamma_n\psi_k} \eta_{n,k}| \\
&\quad + |\nu_{n,k}|/(1 + \gamma\theta) \\
&= |\nu_{n,k} - \operatorname{prox}_{\gamma_n\phi_k}(\nu_{n,k} + \eta_{n,k}) + \operatorname{prox}_{\gamma_n\phi_k} \eta_{n,k}| \\
&\quad + |\nu_{n,k}|/(1 + \gamma\theta).
\end{aligned} \tag{4.41}$$

Consequently, upon setting $\mu = 1 + 1/(\gamma\theta)$, we obtain

$$(\forall n \in \mathbb{N})(\forall k \in \mathbb{L}_2) \quad |\nu_{n,k}| \leq \mu|\nu_{n,k} - \operatorname{prox}_{\gamma_n\phi_k}(\nu_{n,k} + \eta_{n,k}) + \operatorname{prox}_{\gamma_n\phi_k} \eta_{n,k}|. \tag{4.42}$$

In turn,

$$(\forall n \in \mathbb{N}) \quad \sum_{k\in\mathbb{L}_2} |\nu_{n,k}|^2 \leq \mu^2 \sum_{k\in\mathbb{L}_2} |\nu_{n,k} - \operatorname{prox}_{\gamma_n\phi_k}(\nu_{n,k} + \eta_{n,k}) + \operatorname{prox}_{\gamma_n\phi_k} \eta_{n,k}|^2. \tag{4.43}$$

Hence, (4.16) forces $\sum_{k\in\mathbb{L}_2} |\nu_{n,k}|^2 \to 0$, as desired. □

Remark 4.6. An important aspect of Theorem 4.5 is that it provides a *strong* convergence result. Indeed, in general, only weak convergence can be claimed for forward-backward methods [16, 38] (see [3], [4], [16, Remark 5.12], and [25] for explicit constructions in which strong convergence fails). In addition, the standard sufficient conditions for strong convergence in this type of algorithm (see [13, Remark 6.6] and [16, Theorem 3.4(iv)]) are not satisfied in Problem 1.3. Further aspects of the relevance of strong convergence in proximal methods are discussed in [25, 26].

Remark 4.7. Let T be a nonzero bounded linear operator from $\mathcal{H}$ to a real Hilbert space $\mathcal{G}$, let $z \in \mathcal{G}$, and let τ and ω be in $]0, +\infty[$. Specializing Theorem 4.5 to the case when $\Phi\colon x \mapsto \|Tx - z\|^2/2$ and either

$$\mathbb{K} = \varnothing \quad \text{and} \quad (\forall k \in \mathbb{L}) \quad \psi_k = \tau_k|\cdot|^p, \quad \text{where} \quad p \in]1, 2] \ \text{ and } \ \tau_k \in [\tau, +\infty[\,, \tag{4.44}$$

or

$$\mathbb{L} = \varnothing \quad \text{and} \quad (\forall k \in \mathbb{K}) \quad \psi_k = 0 \quad \text{and} \quad \Omega_k = [-\omega_k, \omega_k], \quad \text{where} \quad \omega_k \in [\omega, +\infty[\,, \tag{4.45}$$

yields [16, Corollary 5.19]. If we further impose $\lambda_n \equiv 1$, $\|T\| < 1$, $\gamma_n \equiv 1$, $\alpha_{n,k} \equiv 0$, and $b_n \equiv 0$, we obtain [18, Theorem 3.1].

5. Applications to sparse signal recovery.

5.1. A special case of Problem 1.3. In (1.4), a single observation z of the original signal $\overline{x}$ is available. In certain problems, q such noisy linear observations are available, say $z_i = T_i\overline{x} + v_i$ $(1 \leq i \leq q)$, which leads to the weighted least-squares data fidelity term $x \mapsto \sum_{i=1}^{q} \mu_i \|T_i x - z_i\|^2$; see [12] and the references therein. Furthermore, signal recovery problems are typically accompanied with convex constraints that confine $\overline{x}$ to some closed convex subsets $(S_i)_{1\leq i\leq m}$ of $\mathcal{H}$. The violation of these constraints can be penalized via the cost function $x \mapsto \sum_{i=1}^{m} \vartheta_i d_{S_i}^2(x)$; see [10, 28] and the references therein. On the other hand, power functions are frequently used as cost functions in variational models for determining the coefficients of orthonormal basis decompositions, e.g., [1, 7, 8, 18]. Moreover, we aim at promoting sparsity of a solution $x \in \mathcal{H}$ with respect to $(e_k)_{k\in\mathbb{N}}$ in the sense that, for every k in $\mathbb{K}$, we wish to set to 0 the coefficient $\langle x \mid e_k\rangle$ if it lies in the interval Ω_k. The following formulation is consistent with these considerations.

PROBLEM 5.1. For every $i \in \{1,\ldots,q\}$, let $\mu_i \in]0,+\infty[$, let T_i be a nonzero bounded linear operator from $\mathcal{H}$ to a real Hilbert space $\mathcal{G}_i$, and let $z_i \in \mathcal{G}_i$. For every $i \in \{1,\ldots,m\}$, let $\vartheta_i \in]0,+\infty[$, and let S_i be a nonempty closed and convex subset of $\mathcal{H}$. Furthermore, let $(p_{k,l})_{0\leq l\leq L_k}$ be distinct real numbers in $]1,+\infty[$, let $(\tau_{k,l})_{0\leq l\leq L_k}$ be real numbers in $[0,+\infty[$, and let $l_k \in \{0,\ldots,L_k\}$ satisfy $p_{k,l_k} = \min_{0\leq l\leq L_k} p_{k,l}$, where $(L_k)_{k\in\mathbb{N}}$ is a sequence in $\mathbb{N}$. Finally, let $\mathbb{K} \subset \mathbb{N}$, let $\mathbb{L} = \mathbb{N} \smallsetminus \mathbb{K}$, and let $(\Omega_k)_{k\in\mathbb{K}}$ be a sequence of closed intervals in $\mathbb{R}$. The objective is to

$$\underset{x\in\mathcal{H}}{\text{minimize}} \ \frac{1}{2}\sum_{i=1}^{q} \mu_i \|T_i x - z_i\|^2 + \frac{1}{2}\sum_{i=1}^{m} \vartheta_i d_{S_i}^2(x) + \sum_{k\in\mathbb{N}} \sum_{l=0}^{L_k} \tau_{k,l} |\langle x \mid e_k\rangle|^{p_{k,l}} + \sum_{k\in\mathbb{K}} \sigma_{\Omega_k}(\langle x \mid e_k\rangle), \tag{5.1}$$

under the following assumptions:

(i) $\inf_{k\in\mathbb{L}} \tau_{k,l_k} > 0$,
(ii) $\inf_{k\in\mathbb{L}} p_{k,l_k} > 1$,
(iii) $\sup_{k\in\mathbb{L}} p_{k,l_k} \leq 2$,
(iv) $0 \in \operatorname{int} \bigcap_{k\in\mathbb{K}} \Omega_k$.

PROPOSITION 5.2. *Problem* 5.1 *is a special case of Problem* 1.3.

Proof. First, we observe that (5.1) corresponds to (1.7) where

$$\Phi\colon x \mapsto \frac{1}{2}\sum_{i=1}^{q} \mu_i \|T_i x - z_i\|^2 + \frac{1}{2}\sum_{i=1}^{m} \vartheta_i d_{S_i}^2(x) \quad \text{and} \quad (\forall k \in \mathbb{N}) \quad \psi_k\colon \xi \mapsto \sum_{l=0}^{L_k} \tau_{k,l} |\xi|^{p_{k,l}}. \tag{5.2}$$

Hence, Φ is a finite positive continuous convex function with Fréchet gradient

$$\nabla\Phi\colon x \mapsto \sum_{i=1}^{q} \mu_i T_i^*(T_i x - z_i) + \sum_{i=1}^{m} \vartheta_i (x - P_i x), \tag{5.3}$$

where P_i is the projection operator onto S_i. Therefore, since the operators $(\mathrm{Id} - P_i)_{1\leq i\leq m}$ are nonexpansive, it follows that assumption (i) in Problem 1.3 is satisfied with $1/\beta = \sum_{i=1}^{q} \mu_i \|T_i\|^2 + \sum_{i=1}^{m} \vartheta_i$. Moreover, the functions $(\psi_k)_{k\in\mathbb{N}}$ are in $\Gamma_0(\mathbb{R})$ and satisfy assumptions (ii) and (iii) in Problem 1.3.

Let us now turn to assumption (iv) in Problem 1.3. Fix $\rho \in]0,+\infty[$ and set $\tau = \inf_{k\in\mathbb{L}} \tau_{k,l_k}$, $p = \inf_{k\in\mathbb{L}} p_{k,l_k}$, and $\theta = \tau p(p-1)\min\{1, 1/\rho\}$. Then it follows from (i), (ii), and (iii) that $\theta > 0$ and that

$$
\begin{aligned}
\inf_{k\in\mathbb{L}} \inf_{0<|\xi|\le\rho} \psi_k''(\xi) &= \inf_{k\in\mathbb{L}} \inf_{0<|\xi|\le\rho} \sum_{l=0}^{L_k} \tau_{k,l} p_{k,l}(p_{k,l}-1)|\xi|^{p_{k,l}-2} \\
&\ge \inf_{k\in\mathbb{L}} \tau_{k,l_k} p_{k,l_k}(p_{k,l_k}-1) \inf_{0<\xi\le\rho} \xi^{p_{k,l_k}-2} \\
&\ge \tau p(p-1) \inf_{k\in\mathbb{L}} \inf_{0<\xi\le\rho} \xi^{p_{k,l_k}-2} \\
&\ge \tau p(p-1) \inf_{k\in\mathbb{L}} (1/\rho)^{2-p_{k,l_k}} \\
&\ge \theta,
\end{aligned} \tag{5.4}
$$

which shows that (1.8) is satisfied.

It remains to check assumption (v) in Problem 1.3. To this end, let $\|\cdot\|_{\mathbb{L}}$ denote the standard norm on $\ell^2(\mathbb{L})$, take $\mathsf{x} = (\xi_k)_{k\in\mathbb{L}} \in \ell^2(\mathbb{L})$ such that $\|\mathsf{x}\|_{\mathbb{L}} \ge 1$, and set $(\eta_k)_{k\in\mathbb{L}} = \mathsf{x}/\|\mathsf{x}\|_{\mathbb{L}}$. Then, for every $k \in \mathbb{L}$, $|\eta_k| \le 1$ and, since $p_{k,l_k} \in]1,2]$, we have $|\eta_k|^{p_{k,l_k}} \ge |\eta_k|^2$. Consequently,

$$
\begin{aligned}
\Upsilon_{\mathbb{L}}(\mathsf{x}) &= \sum_{k\in\mathbb{L}} \sum_{l=0}^{L_k} \tau_{k,l}|\xi_k|^{p_{k,l}} \ge \sum_{k\in\mathbb{L}} \tau_{k,l_k}|\xi_k|^{p_{k,l_k}} \\
&\ge \tau \sum_{k\in\mathbb{L}} |\xi_k|^{p_{k,l_k}} = \tau \sum_{k\in\mathbb{L}} \|\mathsf{x}\|_{\mathbb{L}}^{p_{k,l_k}} |\eta_k|^{p_{k,l_k}} \\
&\ge \tau \sum_{k\in\mathbb{L}} \|\mathsf{x}\|_{\mathbb{L}}^{p_{k,l_k}} |\eta_k|^2 = \tau \sum_{k\in\mathbb{L}} \|\mathsf{x}\|_{\mathbb{L}}^{p_{k,l_k}-2} |\xi_k|^2 \\
&\ge \tau \|\mathsf{x}\|_{\mathbb{L}}^{-1} \sum_{k\in\mathbb{L}} |\xi_k|^2 = \tau\|\mathsf{x}\|_{\mathbb{L}}.
\end{aligned} \tag{5.5}
$$

We conclude that $\Upsilon_{\mathbb{L}}(\mathsf{x}) \to +\infty$ as $\|\mathsf{x}\|_{\mathbb{L}} \to +\infty$. □

5.2. First example. Our first example concerns the simulated X-ray fluorescence spectrum $\overline{x}$ displayed in Figure 5.1, which is often used to test restoration methods, e.g., [14, 37]. The measured signal z shown in Figure 5.2 has undergone blurring by the limited resolution of the spectrometer and further corruption by addition of noise. In the underlying Hilbert space $\mathcal{H} = \ell^2(\mathbb{N})$, this process is modeled by $z = T\overline{x} + v$, where $T\colon \mathcal{H} \to \mathcal{H}$ is the operator of convolution with a truncated Gaussian kernel. The noise samples are uncorrelated and drawn from a Gaussian population with mean zero and standard deviation 0.15. The original signal $\overline{x}$ has support $\{0, \dots, N-1\}$ ($N = 1024$), takes on positive values, and possesses a sparse structure. These features can be promoted in Problem 5.1 by letting $(e_k)_{k\in\mathbb{N}}$ be the canonical orthonormal basis of $\mathcal{H}$ and setting $\mathbb{K} = \mathbb{N}$, $\tau_{k,l} \equiv 0$ and

$$
(\forall k \in \mathbb{N}) \quad \Omega_k = \begin{cases}]-\infty, \omega] & \text{if } 0 \le k \le N-1, \\ \mathbb{R} & \text{otherwise,} \end{cases} \tag{5.6}
$$

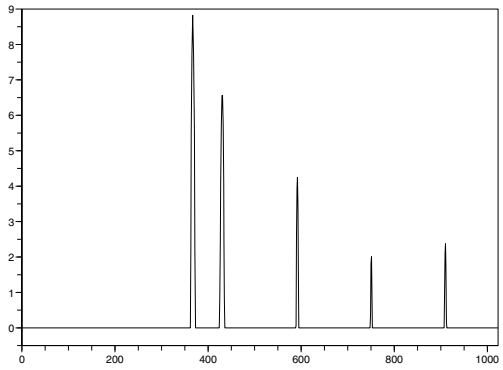


FIG. 5.1. *Original signal—first example.*

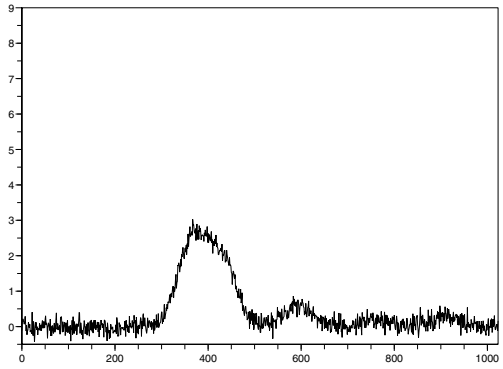


FIG. 5.2. *Degraded signal—first example.*

where the one-sided thresholding level is set to $\omega = 0.01$. On the other hand, using the methodology described in [37], the above information about the noise can be used to construct the constraint sets $S_1 = \big\{x \in \mathcal{H} \;\big|\; \|Tx - z\| \leq \delta_1\big\}$ and $S_2 = \bigcap_{l=1}^{N-1} \big\{x \in \mathcal{H} \;\big|\; |\widehat{Tx}(l/N) - \widehat{z}(l/N)| \leq \delta_2\big\}$, where $\widehat{a}\colon \nu \mapsto \sum_{k=0}^{+\infty} \langle a \mid e_k\rangle \exp(-\imath 2\pi k\nu)$ designates the Fourier transform of $a \in \mathcal{H}$. The bounds δ_1 and δ_2 have been determined so as to guarantee that $\overline{x}$ lies in S_1 and in S_2 with a 99 percent confidence level (see [15] for details). Finally, we set $q = 0$, $m = 2$, and $\vartheta_1 = \vartheta_2 = 1$ in (5.1) (the computation of the projectors P_1 and P_2 required in (5.3) is detailed in [37]). The solution produced by Algorithm 4.3 is shown in Figure 5.3. It is of much better quality than the restorations obtained in [14] and [37] via alternative methods.

5.3. Second example. We provide a wavelet deconvolution example in $\mathcal{H} = \mathbb{L}^2(\mathbb{R})$. The original signal $\overline{x}$ is the classical "bumps" signal [40] displayed in Figure 5.4. The degraded version shown in Figure 5.5 is $z_1 = T_1\overline{x} + v_1$, where T_1 models convolution with a uniform kernel and v_1 is a realization of a zero-mean white Gaussian noise.

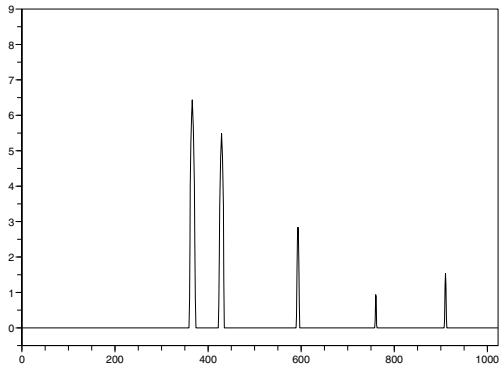


FIG. 5.3. *Signal restored by Algorithm* 4.3*—first example.*

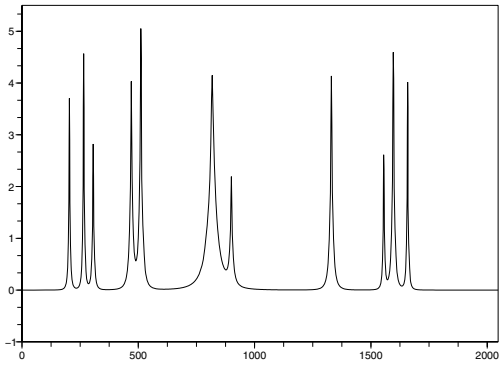


FIG. 5.4. *Original signal—second example.*

The basis $(e_k)_{k\in\mathbb{N}}$ is an orthonormal wavelet symlet basis with eight vanishing moments [17]. Such wavelet bases are known to provide sparse representations for a wide class of signals [22] such as this standard test signal. Note that there exists a strong connection between Problem 5.1 and maximum a posteriori techniques for estimating $\overline{x}$ in the presence of white Gaussian noise. In particular, setting $q = 1$, $m = 0$, $\mathbb{K} = \varnothing$ and $L_k \equiv 0$, and using suitably subband-adapted values of $p_{k,0}$ and $\tau_{k,0}$ amounts to fitting an appropriate generalized Gaussian prior distribution to the wavelet coefficients in each subband [1]. Such a statistical modeling is commonly used in wavelet-based estimation, where values of $p_{k,0}$ close to 2 may provide a good model at coarse resolution levels, whereas values close to 1 should preferably be used at finer resolutions.

The setting of the more general model we adopt here is the following: in Problem 5.1, $\mathbb{K}$ and $\mathbb{L}$ are the index sets of the detail and approximation coefficients [29],

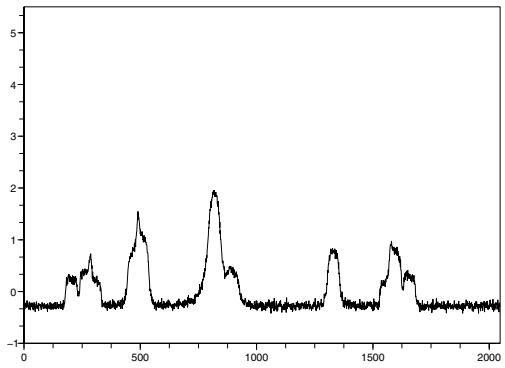


FIG. 5.5. *Degraded signal—second example.*

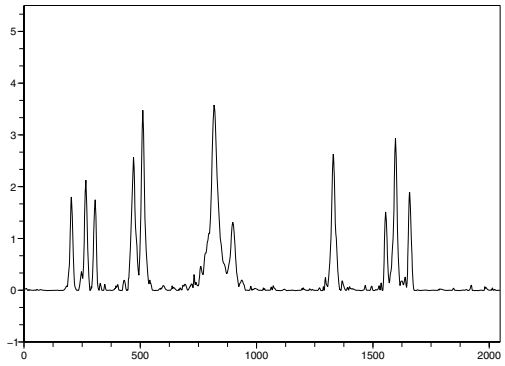


FIG. 5.6. *Signal restored by Algorithm* 4.3*—second example.*

respectively, and

- $(\forall k \in \mathbb{K})$ $\Omega_k = [-0.0023, 0.0023]$, $L_k = 1$, $(p_{k,0}, p_{k,1}) = (2, 4)$, $(\tau_{k,0}, \tau_{k,1}) = (0.0052, 0.0001)$,
- $(\forall k \in \mathbb{L})$ $L_k = 0$, $p_{k,0} = 2$, $\tau_{k,0} = 0.00083$.

For each k, the integer L_k and the exponents $(p_{k,l})_{0 \leq l \leq L_k}$ are imposed, while the set Ω_k and the coefficients $(\tau_{k,l})_{0 \leq l \leq L_k}$ are chosen empirically. In addition, we set $q = 1$, $\mu_1 = 1$, $m = 1$, $\vartheta_1 = 1$, and $S_1 = \big\{x \in \mathcal{H} \;\big|\; x \geq 0\big\}$ (pointwise positivity constraint). The solution x produced by Algorithm 4.3 is shown in Figure 5.6. The estimation error is $\|x - \overline{x}\| = 8.33$. For comparison, the signal $\widetilde{x}$ restored via (1.4) with Algorithm (1.5) is displayed in Figure 5.7. In Problem 5.1, this corresponds to $q = 1$, $m = 0$, $\mathbb{K} = \mathbb{N}$, $\tau_{k,l} \equiv 0$, $\Omega_k \equiv [-2.9, 2.9]$ for the detail coefficients, and $\Omega_k \equiv [-0.0062, 0.0062]$ for the approximation coefficients. This setup yields a worse error of $\|\widetilde{x} - \overline{x}\| = 14.14$ (the sets $(\Omega_k)_{k \in \mathbb{N}}$ have been adjusted so as to mininize this error). The above results have been obtained with a discrete implementation of the wavelet decomposition over four resolution levels using 2048 signal samples [29].

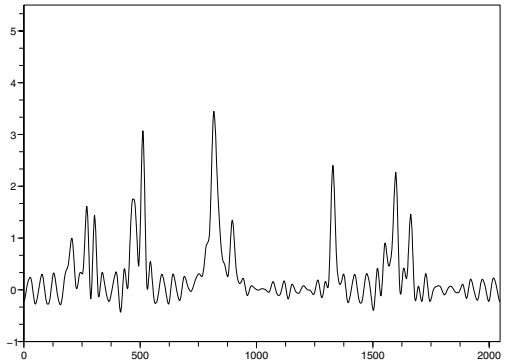


FIG. 5.7. *Signal restored by solving* (1.4)*—second example.*

REFERENCES

[1] A. ANTONIADIS, D. LEPORINI, AND J.-C. PESQUET, *Wavelet thresholding for some classes of non-Gaussian noise*, Statist. Neerlandica, 56 (2002), pp. 434–453.

[2] S. BACCHELLI AND S. PAPI, *Filtered wavelet thresholding methods*, J. Comput. Appl. Math., 164/165 (2004), pp. 39–52.

[3] H. H. BAUSCHKE, J. V. BURKE, F. R. DEUTSCH, H. S. HUNDAL, AND J. D. VANDERWERFF, *A new proximal point iteration that converges weakly but not in norm*, Proc. Amer. Math. Soc., 133 (2005), pp. 1829–1835.

[4] H. H. BAUSCHKE, E. MATOUŠKOVÁ, AND S. REICH, *Projection and proximal point methods: Convergence results and counterexamples*, Nonlinear Anal., 56 (2004), pp. 715–738.

[5] J. BECT, L. BLANC-FÉRAUD, G. AUBERT, AND A. CHAMBOLLE, *A ℓ^1-unified variational framework for image restoration*, in Proceedings of the Eighth European Conference on Computer Vision, Prague, 2004, Lecture Notes in Comput. Sci. 3024, T. Pajdla and J. Matas, eds., Springer-Verlag, New York, 2004, pp. 1–13.

[6] R. E. BRUCK AND S. REICH, *Nonexpansive projections and resolvents of accretive operators in Banach spaces*, Houston J. Math., 3 (1977), pp. 459–470.

[7] A. CHAMBOLLE, R. A. DEVORE, N. Y. LEE, AND B. J. LUCIER, *Nonlinear wavelet image processing: Variational problems, compression, and noise removal through wavelet shrinkage*, IEEE Trans. Image Process., 7 (1998), pp. 319–335.

[8] C. CHAUX, P. L. COMBETTES, J.-C. PESQUET, AND V. R. WAJS, *A variational formulation for frame-based inverse problems*, Inverse Problems, 23 (2007), pp. 1495–1518.

[9] S. CHEN, D. DONOHO, AND M. SAUNDERS, *Atomic decomposition by basis pursuit*, SIAM Rev., 43 (2001), pp. 129–159.

[10] P. L. COMBETTES, *Inconsistent signal feasibility problems: Least-squares solutions in a product space*, IEEE Trans. Signal Process., 42 (1994), pp. 2955–2966.

[11] P. L. COMBETTES, *Convexité et signal*, in Actes du Congrès de Mathématiques Appliquées et Industrielles SMAI'01, Pompadour, France, 2001, pp. 6–16.

[12] P. L. COMBETTES, *A block-iterative surrogate constraint splitting method for quadratic signal recovery*, IEEE Trans. Signal Process., 51 (2003), pp. 1771–1782.

[13] P. L. COMBETTES, *Solving monotone inclusions via compositions of nonexpansive averaged operators*, Optimization, 53 (2004), pp. 475–504.

[14] P. L. COMBETTES AND H. J. TRUSSELL, *Method of successive projections for finding a common point of sets in metric spaces*, J. Optim. Theory Appl., 67 (1990), pp. 487–507.

[15] P. L. COMBETTES AND H. J. TRUSSELL, *The use of noise properties in set theoretic estimation*, IEEE Trans. Signal Process., 39 (1991), pp. 1630–1641.

[16] P. L. COMBETTES AND V. R. WAJS, *Signal recovery by proximal forward-backward splitting*, Multiscale Model. Simul., 4 (2005), pp. 1168–1200.

[17] I. DAUBECHIES, *Ten Lectures on Wavelets*, SIAM, Philadelphia, PA, 1992.

[18] I. Daubechies, M. Defrise, and C. De Mol, *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*, Comm. Pure Appl. Math., 57 (2004), pp. 1413–1457.
[19] I. Daubechies and G. Teschke, *Variational image restoration by means of wavelets: Simultaneous decomposition, deblurring, and denoising*, Appl. Comput. Harmon. Anal., 19 (2005), pp. 1–16.
[20] C. de Mol and M. Defrise, *A note on wavelet-based inversion algorithms*, Contemp. Math., 313 (2002), pp. 85–96.
[21] D. L. Donoho and I. M. Johnstone, *Ideal spatial adaptation via wavelet shrinkage*, Biometrika, 81 (1994), pp. 425–455.
[22] D. L. Donoho and I. M. Johnstone, *Adapting to unknown smoothness via wavelet shrinkage*, J. Amer. Stat. Assoc., 90 (1995), pp. 1200–1224.
[23] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard, *Wavelet shrinkage: Asymptopia?* J. Roy. Statist. Soc. Ser. B, 57 (1995), pp. 301–369.
[24] M. A. T. Figueiredo and R. D. Nowak, *An EM algorithm for wavelet-based image restoration*, IEEE Trans. Image Process., 12 (2003), pp. 906–916.
[25] O. Güler, *On the convergence of the proximal point algorithm for convex minimization*, SIAM J. Control Optim., 29 (1991), pp. 403–419.
[26] O. Güler, *Convergence rate estimates for the gradient differential inclusion*, Optim. Methods Softw., 20 (2005), pp. 729–735.
[27] P. J. Huber, *Robust regression: Asymptotics, conjectures, and Monte Carlo*, Ann. Statist., 1 (1973), pp. 799–821.
[28] T. Kotzer, N. Cohen, and J. Shamir, *A projection-based algorithm for consistent and inconsistent constraints*, SIAM J. Optim., 7 (1997), pp. 527–546.
[29] S. G. Mallat, *A Wavelet Tour of Signal Processing*, 2nd ed., Academic Press, New York, 1999.
[30] J.-J. Moreau, *Fonctions convexes duales et points proximaux dans un espace hilbertien*, C. R. Acad. Sci. Paris, A255 (1962), pp. 2897–2899.
[31] J.-J. Moreau, *Proximité et dualité dans un espace hilbertien*, Bull. Soc. Math. France, 93 (1965), pp. 273–299.
[32] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[33] G. Steidl, J. Weickert, T. Brox, P. Mrázek, and M. Welk, *On the equivalence of soft wavelet shrinkage, total variation diffusion, total variation regularization, and SIDEs*, SIAM J. Numer. Anal., 42 (2004), pp. 686–713.
[34] T. Tao and B. Vidakovic, *Almost everywhere behavior of general wavelet shrinkage operators*, Appl. Comput. Harmon. Anal., 9 (2000), pp. 72–82.
[35] V. N. Temlyakov, *Universal bases and greedy algorithms for anisotropic function classes*, Constr. Approx., 18 (2002), pp. 529–550.
[36] J. A. Tropp, *Just relax: Convex programming methods for identifying sparse signals in noise*, IEEE Trans. Inform. Theory, 52 (2006), pp. 1030–1051.
[37] H. J. Trussell and M. R. Civanlar, *The feasible solution in signal restoration*, IEEE Trans. Acoust., Speech, Signal Process., 32 (1984), pp. 201–212.
[38] P. Tseng, *Applications of a splitting algorithm to decomposition in convex programming and variational inequalities*, SIAM J. Control Optim., 29 (1991), pp. 119–138.
[39] B. Vidakovic, *Nonlinear wavelet shrinkage with Bayes rules and Bayes factors*, J. Amer. Statist. Assoc., 93 (1998), pp. 173–179.
[40] *WaveLab Toolbox*, Stanford University, Palo Alto, CA, http://www-stat.stanford.edu/∼wavelab/.
[41] C. Zălinescu, *Convex Analysis in General Vector Spaces*, World Scientific, River Edge, NJ, 2002.

SIAM J. OPTIM.
Vol. 18, No. 4, pp. 1377–1397

ON MEHROTRA-TYPE PREDICTOR-CORRECTOR ALGORITHMS*

M. SALAHI†, J. PENG‡, AND T. TERLAKY§

Abstract. In this paper we discuss the polynomiality of a feasible version of Mehrotra's predictor-corrector algorithm whose variants have been widely used in several interior point method (IPM)-based optimization packages. A numerical example is given that shows that the adaptive choice of centering parameter and correction terms in this algorithm may lead to small steps being taken in order to keep the iterates in a large neighborhood of the central path, which is important for proving polynomial complexity properties of this method. Motivated by this example, we introduce a safeguard in Mehrotra's algorithm that keeps the iterates in the prescribed neighborhood and allows us to obtain a positive lower bound on the step size. This safeguard strategy is also used when the affine scaling direction performs poorly. We prove that the safeguarded algorithm will terminate after at most $\mathcal{O}(n^2\log(x^0)^T s^0/\epsilon)$ iterations. By modestly modifying the corrector direction, we reduce the iteration complexity to $\mathcal{O}(n\log(x^0)^T s^0/\epsilon)$. To ensure fast asymptotic convergence of the algorithm, we changed Mehrotra's updating scheme of the centering parameter slightly while keeping the safeguard. The new algorithms have the same order of iteration complexity as the safeguarded algorithms but enjoy superlinear convergence as well. Numerical results using the McIPM and LIPSOL software packages are reported.

Key words. linear optimization, predictor-corrector method, interior point methods, Mehrotra-type algorithm, polynomial complexity, superlinear convergence

AMS subject classifications. 90C05, 90C51

DOI. 10.1137/050628787

1. Introduction.

Since Karmarkar's landmark paper [11], the study on interior point methods (IPMs) has become one of the most active research areas in the field of optimization. Many IPMs have been proposed and analyzed [22, 26, 27], and several powerful IPM-based software packages have been developed and successfully applied to numerous applications [1, 5, 29, 30]. Among various variants of IPMs, the so-called predictor-corrector methods have attracted much attention in the IPM community due to its high efficiency and have become the backbones of several optimization packages. It should be mentioned that most implementations of predictor-corrector IPMs adopted a heuristics proposed first by Mehrotra in his remarkable paper [6, 13]. The practical importance of Mehrotra's algorithm motivated us to investigate its theoretical properties. Before going into the details of the algorithm, we briefly review the basics and unique results of IPMs and predictor-corrector IPMs.

We consider primal-dual IPMs for solving the following *linear optimization* (LO) problem

$$\text{(P)} \qquad \min \{c^T x : Ax = b,\ x \geq 0\},$$

*Received by the editors April 7, 2005; accepted for publication (in revised form) June 29, 2007; published electronically December 21, 2007.

http://www.siam.org/journals/siopt/18-4/62878.html

†Department of Mathematics, Faculty of Sciences, The University of Guilan, P.O. Box 1914, Rasht, Iran (salahim@guilan.ac.ir, msalahi@optlab.mcmaster.ca). This author's research was supported by NSERC Discovery grant DG:5-48923, the Canada Research Chair program, and MITACS.

‡Department of Industrial and Enterprise Systems Engineering, University of Illinois, Urbana-Champaign, IL 61801 (pengj@uiuc.edu). This author's research was supported by NSERC Discovery grant DG:249635-02, a PREA award, and MITACS.

§Department of Computing and Software, Advanced Optimization Lab, McMaster University, Hamilton, ON, L8S 4L7, Canada (terlaky@mcmaster.ca). This author's research was supported by NSERC Discovery grant DG:5-48923, the Canada Research Chair program, and MITACS.

where $A \in R^{m\times n}$ satisfies $\text{rank}(A) = m$, $b \in R^m$, $c \in R^n$, and its dual problem

$$\text{(D)} \qquad \max \{b^T y : A^T y + s = c,\ s \geq 0\}.$$

It is common in IPMs theory to assume that both (P) and (D) satisfy the interior point condition (IPC) [22]; i.e., there exists an (x^0, y^0, s^0) such that

$$Ax^0 = b\ x^0 > 0, \quad A^T y^0 + s^0 = c,\ s^0 > 0.$$

Finding optimal solutions of (P) and (D) is equivalent to solving the following system:

$$\begin{aligned} Ax &= b, \quad x \geq 0, \\ A^T y + s &= c, \quad s \geq 0, \\ Xs &= 0, \end{aligned} \tag{1}$$

where $X = \text{diag}\,(x)$. The basic idea of primal-dual IPMs is to replace the third equation in (1) by the parameterized equation $Xs = \mu e$, where e is the all one vector. This leads to the following system:

$$\begin{aligned} Ax &= b, \quad x \geq 0, \\ A^T y + s &= c, \quad s \geq 0, \\ Xs &= \mu e. \end{aligned} \tag{2}$$

If the IPC holds, then for each $\mu > 0$, system (2) has a unique solution. This solution, denoted by $(x(\mu), y(\mu), s(\mu))$, is called the μ-center of the primal-dual pair (P) and (D). The set of μ-centers with all $\mu > 0$ gives *the central path* of (P) and (D) [12, 24]. It has been shown that the limit of the central path (as μ goes to zero) exists and is an optimal solution of (P) and (D) [22].

Applying Newton's method to (2) for a given feasible point (x, y, s) gives the following linear system of equations:

$$\begin{aligned} A\Delta x &= 0, \\ A^T \Delta y + \Delta s &= 0, \\ x\Delta s + s\Delta x &= \mu e - xs, \end{aligned} \tag{3}$$

where $(\Delta x,\ \Delta y,\ \Delta s)$ give the Newton step.

Predictor-corrector algorithms use (3) with different values of μ in the predictor and corrector steps. The predictor-corrector algorithm with best iteration complexity is the Mizuno–Todd–Ye (MTY) algorithm for LO, which operates in two small neighborhoods of the central path [16]. In the predictor step the MTY algorithm uses the so-called primal-dual affine scaling step with $\mu = 0$ in (3) and moves to a slightly larger neighborhood. Then, in the corrector step, it uses $\mu = \mu_g = \frac{x^T s}{n}$, proportional to the duality gap, to bring the iterate towards the central path, back to the smaller neighborhood. In spite of its strong theoretical results for LO and conic linear optimization problems, the algorithm has not been used in developing IPM-based software packages. Several variants of MTY-type predictor-corrector algorithms operating in both small and large neighborhoods have been proposed in the past decade [2, 7, 8, 18, 10, 19, 20, 21, 23], and most of them follow a similar theoretical framework as in [16].

In what follows we describe in detail a feasible[1] version of Mehrotra's original predictor-corrector algorithm that has been widely used in implementations [1, 13, 30]. In the predictor step Mehrotra's algorithm computes the affine scaling search direction, i.e.,

$$\begin{aligned} A\Delta^{\mathrm{a}}x &= 0, \\ A^T\Delta^{\mathrm{a}}y + \Delta^{\mathrm{a}}s &= 0, \\ s\Delta^{\mathrm{a}}x + x\Delta^{\mathrm{a}}x &= -xs; \end{aligned} \tag{4}$$

then it computes the maximum feasible step size that ensures

$$(x + \alpha_{\mathrm{a}}\Delta^{\mathrm{a}}x, s + \alpha_{\mathrm{a}}\Delta^{\mathrm{a}}s) \geq 0.$$

However, the algorithm does not take such a step right away. It is worth mentioning that Mehrotra's original algorithm allows different step sizes in both primal and dual spaces, while here for simplicity of the analysis we consider only the case when they are equal. Mehrotra's algorithm then uses the information from the predictor step to compute the corrector direction that is defined as follows:

$$\begin{aligned} A\Delta x &= 0, \\ A^T\Delta y + \Delta s &= 0, \\ s\Delta x + x\Delta s &= \mu e - xs - \Delta^{\mathrm{a}}x\Delta^{\mathrm{a}}s, \end{aligned} \tag{5}$$

where μ is defined adaptively by

$$\mu = \left(\frac{g_{\mathrm{a}}}{g}\right)^2 \frac{g_{\mathrm{a}}}{n},$$

where $g_{\mathrm{a}} = (x + \alpha_{\mathrm{a}}\Delta^{\mathrm{a}}x)^T(s + \alpha_{\mathrm{a}}\Delta^{\mathrm{a}}s)$ and $g = x^Ts$. Since $(\Delta^{\mathrm{a}}x)^T\Delta^{\mathrm{a}}s = 0$, the previous relation implies

$$\mu = (1 - \alpha_{\mathrm{a}})^3\mu_g. \tag{6}$$

From (6) it is obvious that if only a small step in the affine scaling direction can be made, then we improve only the centrality of the iterate.

Finally, Mehrotra's algorithm makes a step in the $(\Delta x, \Delta y, \Delta s)$ direction by an appropriate step size, and let us denote the new iterate by

$$x(\alpha) := x + \alpha\Delta x, \quad y(\alpha) := y + \alpha\Delta y, \quad s(\alpha) := s + \alpha\Delta s.$$

We note that several variants of the previous algorithm have been well studied in the literature. For example, Mehrotra proposed an infeasible second order predictor-corrector IPM [14] based on a similar power series extension of Monteiro, Adler, and Resende [17]. In his infeasible variant, Mehrotra combined the adaptive scheme with a safeguard technique to stabilize the convergence of the algorithm. Zhang and Zhang [28] have analyzed this second order algorithm without using the adaptive update of the centrality parameter. Jarre and Wechs [9] have suggested generating several corrector directions first and then using the generated directions to construct a new

[1] The original Mehrotra algorithm is an infeasible algorithm. However, the self-dual embedding model [22] can be used to construct a slightly bigger LO problem that has an obvious starting point on the central path.

search direction along which a step can be taken. Gondzio [6] proposed using multiple centrality steps to bring the iterates back to the vicinity of the central path. Significant improvements have been reported for solving several challenge NETLIB test problems and problems arising from real applications. In a recent work [4], Gondzio further combined the idea of multiple centering with a symmetric neighborhood to avoid potential ill behaviors of Mehrotra's predictor-corrector algorithm. More recently, Mehrotra and Li [15] considered a Krylov subspace-based predictor-corrector method and established its global convergence. Promising numerical results are reported as well.

Different from the above-mentioned results, in this paper we first explore the potential flaws in the feasible version of Mehrotra's original algorithm. By a numerical example we show that Mehrotra's algorithm may result in very small steps in order to keep the iterate in a certain neighborhood of the central path, which is essential to prove the polynomiality of the algorithm. To avoid such a trap, we propose incorporating a safeguard in the algorithm so that we can guarantee a positive lower bound for the step size and subsequently the polynomial complexity. Further, to ensure the superlinear convergence of the algorithm we changed the updating scheme of the centering parameter so that the new scheme preserves the same iteration complexity with stronger asymptotic convergence results. It is worthwhile mentioning that our simple safeguard strategy is different from the most recent results by Colombo and Gondzio [4] and Mehrotra and Li [15], where they employ multiple centering with symmetric neighborhood and Krylov subspace-based corrections, respectively. Most recently, in [3] the author also provided another example that shows that the second order variant of Mehrotra's predictor-corrector algorithms may fail to converge to an optimal solution. However, there have not been any numerical experiments.

The rest of the paper is organized as follows. First, in section 2, we present a numerical example that motivates the introduction of a safeguard in Mehrotra's algorithm. Then, in section 3, we present the safeguard-based algorithm and establish its worst case iteration complexity. For readability of the paper we moved some technical lemmas that are used in section 3 to the appendix. In section 4, we further modify the algorithm of section 3 and discuss its iteration complexity. In section 5, Mehrotra's updating scheme of the centering parameter is slightly modified to ensure the superlinear convergence of both algorithms in sections 3 and 4. Some illustrative numerical results using the NETLIB and Kennington test problems are reported in section 6, and finally we conclude the paper by few remarks in section 7.

Conventions. Throughout the paper $\|\cdot\|$ denotes the 2-norm of vectors. We denote by $\mathcal{I}$ the index set $\{1, 2, \ldots, n\}$. For any two vectors x and s, xs denotes the componentwise product of the two vectors, and e denotes the vector with all components equal to one. For simplicity of notation we remove the iteration index in the coming sections. We also use the notation

$$\mathcal{I}_+ = \{i \in \mathcal{I} \mid \Delta x_i^a \Delta s_i^a > 0\}, \quad \mathcal{I}_- = \{i \in \mathcal{I} \mid \Delta x_i^a \Delta s_i^a < 0\},$$

$$\mathcal{F} = \{(x, y, s) \in R^n \times R^m \times R^n \mid (x, s) \geq 0,\ Ax = b,\ A^T y + s = c\},$$

and

$$\mathcal{F}^0 = \{(x, y, s) \in R^n \times R^m \times R^n \mid (x, s) > 0,\ Ax = b,\ A^T y + s = c\}.$$

2. Motivation. In this section first we introduce the neighborhood of the central path in which the algorithms operate. Then we give a numerical example showing that using the strategy described in the introduction might force the algorithm to make very small steps to keep the iterate in a certain neighborhood of the central path, which further implies the algorithm needs to take many iterations to convergence. The example indicates that Mehrotra's adaptive updating scheme of the centering parameter has to be combined with certain safeguards to get a warranted step size at each iteration.

Most efficient IPM solvers work in the negative infinity norm neighborhood defined by

$$\mathcal{N}_\infty^-(\gamma) := \{(x,y,s) \in \mathcal{F}^0 : \ x_i s_i \geq \gamma \mu_g \ \forall i = 1,\dots,n\}, \tag{7}$$

where $\gamma \in (0,1)$ is a constant independent of n and $\mu_g = \frac{x^T s}{n}$. In this paper, we consider algorithms that are working in $\mathcal{N}_\infty^-(\gamma)$ (called the large neighborhood).

Let us consider the following simple LO:

$$\begin{aligned} \min \quad & -x_2 \\ \text{s.t. } & 0 \leq x_1 \leq 1, \\ & 0 \leq x_2 \leq 1 + \delta x_1, \end{aligned} \tag{8}$$

where $\delta = 0.1$. Let the algorithm start with the following feasible points in the neighborhood $\mathcal{N}_\infty^-(0.1)$:

$$x^0 = (0.03; 0.9), \quad s^0 = (6.8; 1; 7; 2), \quad y^0 = (-7, -2).$$

For the given starting point, if we use identical step sizes for both primal and dual problems, in the third iteration the maximum step size in the predictor step will be $\alpha_{\mathrm{a}} = 0.96$, while the maximum step size in the corrector step is $\mathcal{O}(10^{-4})$, and this value is getting worse for later iterations. To explain what we observed, let us examine the constraints

$$x_i(\alpha) s_i(\alpha) \geq \gamma \mu_g(\alpha) \quad \forall i \in \mathcal{I} \tag{9}$$

for $\gamma = 0.1$ that keeps the next iterate in the $\mathcal{N}_\infty^-(0.1)$ neighborhood, where

$$\mu_g(\alpha) = \frac{x(\alpha)^T s(\alpha)}{n} = \left(1 - \alpha + \alpha \frac{\mu}{\mu_g}\right) \mu_g. \tag{10}$$

By expanding inequality (9) and reordering one has

$$(1-\alpha) x_i s_i + \alpha(1-0.1)\mu - \alpha \Delta x_i^{\mathrm{a}} \Delta s_i^{\mathrm{a}} + \alpha^2 \Delta x_i \Delta s_i \geq \gamma(1-\alpha)\mu_g \quad \forall i \in \mathcal{I}.$$

Note that for the given staring point, $x_1 s_1 - 0.1\mu_g$ is a very small nonnegative number, while $\Delta x_1^{\mathrm{a}} \Delta s_1^{\mathrm{a}}$ and $\Delta x_1 \Delta s_1$ are both negative numbers whose absolute values are dominated by $x_1 s_1$, and finally $\mu = \mathcal{O}(10^{-5})$ due to a big affine scaling step size. Incorporating all these information into (9) implies that the algorithm requires a very small step to satisfy (9). This phenomenon might be the result of the following:

- There is an aggressive update of centering parameter μ using (6).
- There is usage of the correction terms in the corrector system of equations.

To resolve these difficulties, we propose the following remedies:

- Use a fixed fraction of μ_g, for example, $\mu = \frac{\mu_g}{10}$, rather than an adaptive update.
- Cut the maximum step size in the predictor step if it is above a certain threshold. This might prevent the algorithm from having an aggressive update.
- Modify the correction terms in the corrector system of equations.

For this specific example, these ideas help us to solve the difficulty that might arise. However, in general modifying the second order correction terms may not be as effective as using a simple large update of the centering parameter.

These observations motivate us to introduce a safeguard strategy that will help us to have control on the minimal warranted step size from the theoretical and practical points of view. In our safeguard we simply use a fixed fraction of μ_g as the μ value. It is worthwhile mentioning that when the affine scaling step size is very small, for example, when $\alpha_{\rm a} < 0.1$, which implies marginal reduction of the complementarity gap, we also employ the same large update safeguard.

3. A safeguard-based algorithm. In this section we first discuss the step size estimation of the algorithm and then outline the safeguard-based algorithm. Finally, we establish its worst case iteration complexity.

The following technical lemma will be used in the next theorem, which estimates the maximum step size in the corrector step.

LEMMA 3.1. *Suppose the current iterate is* $(x, y, s) \in \mathcal{N}_\infty^-(\gamma)$*, and let* $(\Delta x, \Delta y, \Delta s)$ *be the solution of* (5)*, where* $\mu \geq 0$*. Then we have*

$$||\Delta x \Delta s|| \leq 2^{\frac{-3}{2}} \left(\frac{1}{\gamma} \left(\frac{\mu}{\mu_g} \right)^2 - \left(2 - \frac{1}{2\gamma} \right) \frac{\mu}{\mu_g} + \frac{17\gamma + n}{16\gamma} \right) n\mu_g.$$

Proof. If we multiply the third equation of (5) by $(XS)^{-\frac{1}{2}}$, then by Lemma 5.3 of [26] we have

$$\begin{aligned} ||\Delta x \Delta s|| &\leq 2^{\frac{-3}{2}} ||\mu (XSe)^{-\frac{1}{2}} - (XSe)^{\frac{1}{2}} - (XS)^{-\frac{1}{2}} \Delta x^{\rm a} \Delta s^{\rm a}||^2 \\ &= 2^{\frac{-3}{2}} \left(\mu^2 \sum_{i \in \mathcal{I}} \frac{1}{x_i s_i} + \sum_{i \in \mathcal{I}} x_i s_i + \sum_{i \in \mathcal{I}} \frac{(\Delta x_i^{\rm a} \Delta s_i^{\rm a})^2}{x_i s_i} - 2n\mu - 2\mu \sum_{i \in \mathcal{I}} \frac{\Delta x_i^{\rm a} \Delta s_i^{\rm a}}{x_i s_i} \right) \\ &\leq 2^{\frac{-3}{2}} \left(\frac{n\mu^2}{\gamma \mu_g} + n\mu_g + \frac{n\mu_g}{16} + \frac{n^2 \mu_g}{16\gamma} - 2n\mu + \frac{n\mu}{2\gamma} \right), \end{aligned}$$

where the last inequality follows from Lemma A.2 and the assumption that the previous iterate is in $\mathcal{N}_\infty^-(\gamma)$. By reordering and factorizing we get the statement of the lemma. ∎

Motivated from the computational practice, we use $\mu = \frac{\beta}{1-\beta}\mu_g$ as the value of safeguard, where $\gamma \leq \beta < \frac{1}{3}$. This is due to the fact that using $\mu = \frac{\gamma}{1-\gamma}\mu_g$ for small values of γ might imply an aggressive update of the barrier parameter. The following corollary, which follows from Lemma 3.1, gives an explicit upper bound for this specific value of μ.

COROLLARY 3.2. *If* $\mu = \frac{\beta}{1-\beta}\mu_g$*, where* $\gamma \leq \beta < \frac{1}{3}$ *and* $\gamma \in (0, \frac{1}{3})$*, then*

$$||\Delta x \Delta s|| \leq \frac{1}{2\gamma} n^2 \mu_g.$$

THEOREM 3.3. *Suppose the current iterate is* $(x, y, s) \in \mathcal{N}_\infty^-(\gamma)$*, where* $\gamma \in (0, \frac{1}{3})$*,*

and let $(\Delta x, \Delta y, \Delta s)$ *be the solution of* (5) *with*

$$\mu = \frac{\beta}{1-\beta}\mu_g,$$

where $\gamma \le \beta < \frac{1}{3}$. *Then the maximum step size* α_c, *which keeps* $(x(\alpha_c), y(\alpha_c), s(\alpha_c))$ *in* $\mathcal{N}_\infty^-(\gamma)$, *satisfies*

$$\alpha_c \ge \frac{3\gamma^2}{2n^2}.$$

Proof. The goal is to find the maximum nonnegative α for which the relation (9) holds. To do so, first let use define

$$t = \max_{i \in \mathcal{I}_+} \left\{ \frac{\Delta x_i^a \Delta s_i^a}{x_i s_i} \right\}. \tag{11}$$

Since $(\Delta x^a)^T \Delta s^a = 0$, then $\mathcal{I}_+ \neq \emptyset$. Now it is sufficient to prove (9) for $\Delta x_i^a \Delta s_i^a > 0$. To do so, we have

$$\begin{aligned} x_i(\alpha)s_i(\alpha) &= x_i s_i + \alpha(\mu - x_i s_i - \Delta x^a{}_i \Delta s^a{}_i) + \alpha^2 \Delta x_i \Delta s_i \\ &\ge (1-\alpha)x_i s_i + \alpha\mu - \alpha t x_i s_i - \frac{\alpha^2 n^2 \mu_g}{2\gamma} \\ &= (1-(1+t)\alpha)x_i s_i + \alpha\mu - \frac{\alpha^2 n^2 \mu_g}{2\gamma}, \end{aligned}$$

where the first inequality follows from $\alpha \ge 0$, (11), and Corollary 3.2. Moreover, from Lemma A.1 we have that $t \le \frac{1}{4}$, which implies $\frac{1}{1+t} \ge \frac{4}{5}$. Thus we further deduce that for $\alpha \in [0, \frac{4}{5}]$, we have

$$x_i(\alpha)s_i(\alpha) \ge (1-(1+t)\alpha)\gamma\mu_g + \alpha\mu - \frac{\alpha^2 n^2 \mu_g}{2\gamma}.$$

Now using (10), the next iterate belongs to $\mathcal{N}_\infty^-(\gamma)$, provided

$$(1-(1+t)\alpha)\gamma\mu_g + \alpha\mu - \frac{\alpha^2 n^2 \mu_g}{2\gamma} \ge \gamma\left(1 - \alpha + \alpha\frac{\mu}{\mu_g}\right)\mu_g,$$

which is equivalent to

$$(1-\gamma)\mu - \gamma t \mu_g \ge \frac{\alpha n^2 \mu_g}{2\gamma}. \tag{12}$$

Using Lemma A.1 and the definition of μ one has

$$(1-\gamma)\mu - \gamma t\mu_g \ge \frac{(1-\gamma)\beta}{(1-\beta)}\mu_g - \frac{\gamma\mu_g}{4} \ge \frac{3\gamma\mu_g}{4}.$$

Therefore, inequality (12) holds if

$$\frac{3\gamma\mu_g}{4} \ge \frac{\alpha n^2 \mu_g}{2\gamma}.$$

This inequality definitely holds for $\alpha = \frac{3\gamma^2}{2n^2}$. Now we can conclude that

$$\alpha_c \geq \min\left(\frac{4}{5}, \frac{3\gamma^2}{2n^2}\right) = \frac{3\gamma^2}{2n^2}. \qquad \square$$

We remind the readers that we use this safeguard when the affine scaling performs poorly, for example, when $\alpha_{\rm a} < 0.1$.

Now after all the previous discussions we may outline our new safeguard-based algorithm as follows.

ALGORITHM 1

Input:
A proximity parameters $\gamma \in (0, \frac{1}{3})$;
a safeguard parameter $\beta \in [\gamma, \frac{1}{3})$;
an accuracy parameter $\epsilon > 0$;
$(x^0, y^0, s^0) \in \mathcal{N}_\infty^-(\gamma)$.

begin
while $x^T s \geq \epsilon$ **do**
begin
Predictor Step Solve (4) and compute the maximum step size $\alpha_{\rm a}$ such that $(x(\alpha_{\rm a}), y(\alpha_{\rm a}), s(\alpha_{\rm a})) \in \mathcal{F}$;
end

begin
Corrector step
If $\alpha_{\rm a} \geq 0.1$, **then** solve (5) with $\mu = (1 - \alpha_{\rm a})^3 \mu_g$ and compute the maximum step size $\alpha_{\rm c}$ such that $(x(\alpha_{\rm c}), y(\alpha_{\rm c}), s(\alpha_{\rm c})) \in \mathcal{N}_\infty^-(\gamma)$;
If $\alpha_{\rm c} < \frac{3\gamma^2}{2n^2}$, **then** solve (5) with $\mu = \frac{\beta}{1-\beta}\mu_g$ and compute the maximum step size $\alpha_{\rm c}$ such that $(x(\alpha_{\rm c}), y(\alpha_{\rm c}), s(\alpha_{\rm c})) \in \mathcal{N}_\infty^-(\gamma)$;
end
else
Solve (5) with $\mu = \frac{\beta}{1-\beta}\mu_g$ and compute the maximum step size $\alpha_{\rm c}$ such that $(x(\alpha_{\rm c}), y(\alpha_{\rm c}), s(\alpha_{\rm c})) \in \mathcal{N}_\infty^-(\gamma)$;
end Set $(x(\alpha_{\rm c}), y(\alpha_{\rm c}), s(\alpha_{\rm c})) = (x + \alpha_{\rm c}\Delta x, y + \alpha_{\rm c}\Delta y, s + \alpha_{\rm c}\Delta s)$.
end
end

Remark 3.4. By using an identical step size for both the primal and the dual problems, comparing with the Mehrotra's algorithm, our new algorithm requires at most an extra backsolve to make a better step.

The following theorem gives an upper bound for the maximum number of iterations in which Algorithm 1 stops with an ϵ-approximate solution.

THEOREM 3.5. *Algorithm* 1 *stops after at most*

$$O\left(n^2 \log \frac{(x^0)^T s^0}{\epsilon}\right)$$

iterations with a solution for which $x^T s \leq \epsilon$.

Proof. If $\alpha_a < 0.1$ or $\alpha_c < \frac{3\gamma^2}{2n^2}$, then the algorithm uses the safeguard strategy. It follows from Theorem 3.3 and (10) that

$$\mu_g(\alpha) \leq \left(1 - \frac{3\gamma^2(1-2\beta)}{2(1-\beta)n^2}\right)\mu_g.$$

If $\alpha_a \geq 0.1$ and $\alpha_c \geq \frac{3\gamma^2}{2n^2}$, then the algorithm uses Mehrotra's updating strategy, which further implies that

$$\mu_g(\alpha) \leq \left(1 - \frac{2\gamma^2}{5n^2}\right)\mu_g,$$

which completes the proof conforming to [26, Theorem 3.2]. ∎

4. A modified version of Algorithm 1. In this section we propose a slightly modified version of Algorithm 1 (Algorithm 2) that enjoys much better iteration complexity than Algorithm 1 and also is computationally more appealing. The improvement in the iteration complexity is the result of the following proposition and modified corrector step that allow us to strengthen the bound in Lemma 3.1.

PROPOSITION 4.1. *For all $i \in \mathcal{I}_-$ one has*

$$-\Delta x_i^a \Delta s_i^a \leq \frac{1}{\alpha_a}\left(\frac{1}{\alpha_a} - 1\right) x_i s_i. \tag{13}$$

Proof. For the maximum step size in the predictor step, α_a, one has $x_i(\alpha_a)s_i(\alpha_a) \geq 0,\ i = 1, \ldots, n$. This is equivalent to $(1-\alpha_a)x_i s_i + \alpha_a^2 \Delta x_i^a \Delta s_i^a \geq 0,\ i = 1, \ldots, n$, which the statement of the proposition follows. ∎

The motivation for modifying the Newton system in the corrector step is the following observation. If the maximal feasible step size for the affine scaling direction is reasonably large, then the classical corrector direction should also be a good choice. However, if the maximal feasible step size for the affine scaling search direction is very small, then we should possibly try to bring the iterate back to the vicinity of the central path. In such a case, the second order correction terms in system (5) might not be a good choice since they might lead to a search direction moving towards the boundary of the feasible region. Therefore, we propose changing the second order correction terms in the corrector step proportional to the affine scaling step size when it does not perform good, for example, when $\alpha_a < 0.1$.

The new corrector system of equations when $\alpha_a < 0.1$, which by using Proposition 4.1 enables us to improve on the iteration complexity of Algorithm 1, is

$$\begin{aligned} A\Delta x &= 0, \\ A^T\Delta y + \Delta s &= 0, \\ s\Delta x + x\Delta s &= \mu e - xs - \alpha_a \Delta x^a \Delta s^a, \end{aligned} \tag{14}$$

where the centering parameter μ is defined as in the previous section. Changing the corrector system of equation to (14) when $\alpha_a < 0.1$ helps us avoid the potential ill behaviors of Mehrotra's original algorithm without sacrificing its practical efficiency (see section 6). Thus, in what follows we consider this variant for further analysis.

Analogous to Lemma 3.1 we have the following bound for $\|\Delta x \Delta s\|$ when $\alpha_a \in (0, 0.1)$.

LEMMA 4.2. *Suppose the current iterate is $(x, y, s) \in \mathcal{N}_\infty^-(\gamma)$, where $\alpha_a \in (0, 0.1)$, and let $(\Delta x, \Delta y, \Delta s)$ be the solution of* (14). *Then we have*

$$||\Delta x \Delta s|| \leq 2^{\frac{-3}{2}}\left(\frac{1}{\gamma}\left(\frac{\mu}{\mu_g}\right)^2 - \left(2 - \frac{\alpha_a}{2\gamma}\right)\frac{\mu}{\mu_g} + \frac{20 - 4\alpha_a + \alpha_a^2}{16}\right) n\mu_g.$$

Proof. Since $(\Delta x^a)^T \Delta s^a = 0$, both $\mathcal{I}_+$ and $\mathcal{I}_-$ are nonempty. If we multiply the third equation of (14) by $(XS)^{-\frac{1}{2}}$, then by Lemma 5.3 of [26] we have

$$\begin{aligned}
\|\Delta x \Delta s\| &\leq 2^{\frac{-3}{2}} ||\mu(XSe)^{-\frac{1}{2}} - (XSe)^{\frac{1}{2}} - \alpha_{\rm a}(XS)^{-\frac{1}{2}} \Delta x^a \Delta s^a||^2 \\
&= 2^{\frac{-3}{2}} \left(\mu^2 \sum_{i\in\mathcal{I}} \frac{1}{x_i s_i} + x^T s + \alpha_{\rm a}^2 \sum_{i\in\mathcal{I}} \frac{(\Delta x_i^{\rm a} \Delta s_i^{\rm a})^2}{x_i s_i} - 2n\mu - 2\alpha_{\rm a}\mu \sum_{i\in\mathcal{I}} \frac{\Delta x_i^{\rm a} \Delta s_i^{\rm a}}{x_i s_i} \right) \\
&\leq 2^{\frac{-3}{2}} \left(\frac{n\mu^2}{\gamma\mu_g} + n\mu_g + \frac{\alpha_{\rm a}^2 n\mu_g}{16} - (1-\alpha_{\rm a}) \sum_{i\in\mathcal{I}_-} \Delta x_i^{\rm a} \Delta s_i^{\rm a} - 2n\mu + \frac{\alpha_{\rm a} n\mu}{2\gamma} \right) \\
&\leq 2^{\frac{-3}{2}} \left(\frac{1}{\gamma} \left(\frac{\mu}{\mu_g} \right)^2 + 1 + \frac{\alpha_{\rm a}^2}{16} + \frac{(1-\alpha_{\rm a})}{4} - 2\frac{\mu}{\mu_g} + \frac{\alpha_{\rm a}}{2\gamma} \frac{\mu}{\mu_g} \right) n\mu_g \\
&= 2^{\frac{-3}{2}} \left(\frac{1}{\gamma} \left(\frac{\mu}{\mu_g} \right)^2 - \left(2 - \frac{\alpha_{\rm a}}{2\gamma} \right) \frac{\mu}{\mu_g} + \frac{20 - 4\alpha_{\rm a} + \alpha_{\rm a}^2}{16} \right) n\mu_g,
\end{aligned}$$

where the second inequality follows from (13), Lemmas A.1 and A.2, and the assumption that the previous iterate is in $\mathcal{N}_\infty^-(\gamma)$. The third inequality also follows from Lemma A.2. ∎

The following corollary gives an explicit upper bound for a specific μ.

COROLLARY 4.3. *Let* $\mu = \frac{\beta}{1-\beta}\mu_g$, *where* $\beta \in [\gamma, \frac{1}{3})$, $\gamma \in (0, \frac{1}{3})$, *and* $\alpha_{\rm a} \in (0, 0.1)$; *then*

$$||\Delta x \Delta s|| \leq \frac{\beta}{\sqrt{2}\gamma(1-\beta)} n\mu_g.$$

In the following theorem we estimate the maximum step size in the corrector step of the modified algorithm defined by (14) for $\alpha_{\rm a} \in (0, .01)$.

THEOREM 4.4. *Suppose the current iterate is* $(x, y, s) \in \mathcal{N}_\infty^-(\gamma)$, *where* $\gamma \in (0, \frac{1}{3})$, $\beta \in [\gamma, \frac{1}{3})$, *and* $\alpha_{\rm a} \in (0, 0.1)$, *and* $(\Delta x, \Delta y, \Delta s)$ *is the solution of* (14) *with* $\mu = \frac{\beta}{1-\beta}\mu_g$. *Then the maximum step size* $\alpha_{\rm c}$, *such that* $(x(\alpha_{\rm c}), y(\alpha_{\rm c}), s(\alpha_{\rm c})) \in \mathcal{N}_\infty^-(\gamma)$, *satisfies*

$$\alpha_{\rm c} \geq \frac{39\sqrt{2}\gamma(1-\gamma)}{40n}.$$

Proof. We need to estimate the maximum nonnegative α for which (9) holds. We know that $(\Delta x^{\rm a})^T \Delta s^{\rm a} = 0$; then $\mathcal{I}_+ \neq \emptyset$. It suffices to consider only the case when $\Delta x_i^{\rm a} \Delta s_i^{\rm a} > 0$. Therefore, we have

$$\begin{aligned}
x_i(\alpha) s_i(\alpha) &= x_i s_i + \alpha(\mu - x_i s_i - \alpha_{\rm a} \Delta x^{\rm a}{}_i \Delta s^{\rm a}{}_i) + \alpha^2 \Delta x_i \Delta s_i \\
&\geq (1-\alpha) x_i s_i + \alpha\mu - \alpha\alpha_{\rm a} t x_i s_i - \frac{\beta}{\sqrt{2}\gamma(1-\beta)} \alpha^2 n\mu_g \\
&= (1 - \alpha(1+\alpha_{\rm a} t))\, x_i s_i + \alpha\mu - \frac{\beta}{\sqrt{2}\gamma(1-\beta)} \alpha^2 n\mu_g,
\end{aligned}$$

where the first inequality follows from α being nonnegative, (11), and Corollary 4.3. Moreover, from Lemma A.1 we have that $t \leq \frac{1}{4}$, which implies $\frac{1}{1+\alpha_{\rm a} t} \geq \frac{4}{5}$. Thus we further deduce that for $\alpha \in [0, \frac{4}{5}]$, we have

$$x_i(\alpha) s_i(\alpha) \geq (1 - (1+\alpha_{\rm a} t)\alpha)\gamma\mu_g + \alpha\mu - \frac{\beta}{\sqrt{2}\gamma(1-\beta)} \alpha^2 n\mu_g.$$

By using (10), the new iterate is in $\mathcal{N}_\infty^-(\gamma)$ whenever

$$\gamma\left(1-\alpha(1+\alpha_{\mathrm{a}}t)\right)+\frac{\mu}{\mu_g}\alpha-\frac{\beta}{\sqrt{2}\gamma(1-\beta)}\alpha^2 n\geq\gamma\left(1-\alpha+\frac{\mu}{\mu_g}\alpha\right).$$

Analogous to Theorem 3.3, one can easily verify that this inequality holds for

$$\alpha=\frac{39\sqrt{2}\gamma(1-\gamma)}{40n}.$$

Therefore, we have

$$\alpha_c\geq\min\left(\frac{4}{5},\frac{39\sqrt{2}\gamma(1-\gamma)}{40n}\right)=\frac{39\sqrt{2}\gamma(1-\gamma)}{40n}:=\hat{\alpha}_c. \qquad \square$$

Now we can outline Algorithm 2 as follows.

ALGORITHM 2

Input:
A proximity parameters $\gamma\in(0,\frac{1}{3})$;
a safeguard parameter $\beta\in[\gamma,\frac{1}{3})$;
an accuracy parameter $\epsilon>0$;
$(x^0,y^0,s^0)\in\mathcal{N}_\infty^-(\gamma)$.

begin
while $x^Ts\geq\epsilon$ **do**
begin
Predictor Step
Solve (4) and compute the maximum step size α_{a} such that $(x(\alpha_{\mathrm{a}}),y(\alpha_{\mathrm{a}}),s(\alpha_{\mathrm{a}}))\in\mathcal{F}$;
end

begin
Corrector step
If $\alpha_{\mathrm{a}}\geq 0.1$, **then** solve (5) with $\mu=(1-\alpha_{\mathrm{a}})^3\mu_g$ and compute the maximum step size α_{c} such that $(x(\alpha_{\mathrm{c}}),y(\alpha_{\mathrm{c}}),s(\alpha_{\mathrm{c}}))\in\mathcal{N}_\infty^-(\gamma)$;
If $\alpha_{\mathrm{c}}<\hat{\alpha}_c$, **then** solve (5) with $\mu=\frac{\beta}{1-\beta}\mu_g$ and compute the maximum step size α_{c} such that $(x(\alpha_{\mathrm{c}}),y(\alpha_{\mathrm{c}}),s(\alpha_{\mathrm{c}}))\in\mathcal{N}_\infty^-(\gamma)$;
end
else
Solve (14) with $\mu=\frac{\beta}{1-\beta}\mu_g$ and compute the maximum step size α_{c} such that $(x(\alpha_{\mathrm{c}}),y(\alpha_{\mathrm{c}}),s(\alpha_{\mathrm{c}}))\in\mathcal{N}_\infty^-(\gamma)$;
end
Set $(x(\alpha_{\mathrm{c}}),y(\alpha_{\mathrm{c}}),s(\alpha_{\mathrm{c}}))=(x+\alpha_{\mathrm{c}}\Delta x,y+\alpha_{\mathrm{c}}\Delta y,s+\alpha_{\mathrm{c}}\Delta s)$.
end
end

The following theorem gives an upper bound for the number of iterations in which Algorithm 2 stops with an ϵ-approximate solution.

THEOREM 4.5. *Algorithm* 2*, the modified version of Algorithm* 1*, stops after at most*

$$\mathcal{O}\left(n \log \frac{(x^0)^T s^0}{\epsilon}\right)$$

iterations with a solution for which $x^T s \leq \epsilon$.

Proof. If $\alpha_{\rm a} < 0.1$ or $\alpha_{\rm c} < \hat{\alpha}_c$, then the algorithm uses the safeguard strategy. Then by (10) and Theorem 4.4 one has

$$\mu_g(\alpha) \leq \left(1 - \frac{39\sqrt{2}\gamma(1-\gamma)(1-2\beta)}{40(1-\beta)n}\right)\mu_g.$$

If $\alpha_{\rm a} \geq 0.1$ and $\alpha_{\rm c} \geq \hat{\alpha}_c$, then the algorithm uses Mehrotra's updating strategy, which further implies that

$$\mu_g(\alpha) \leq \left(1 - \frac{37\gamma(1-\gamma)}{100n}\right)\mu_g,$$

which completes the proof conforming to Theorem 3.2 of [26]. ∎

5. Superlinear convergence. In this section we analyze the asymptotic behavior of the previous algorithms using a modification of the centering parameter μ rather than using (6) due to the following observations.

We note that by Theorem 7.4 of [26] for $(x, s) \in \mathcal{N}_\infty^-(\gamma)$ the relations

$$|\Delta x_i^{\rm a} \Delta s_i^{\rm a}| \leq \mathcal{O}(\mu_g^2), \quad i = 1, \ldots, n, \tag{15}$$

hold. This further implies that $\alpha_{\rm a} \geq 1 - \mathcal{O}(\mu_g)$. Now, for the asymptotic case, one has to estimate α that satisfies the following inequalities for each $i \in \mathcal{I}$:

$$(1-\alpha)x_i^k s_i^k + (1-\gamma)\alpha\mu - \alpha\Delta x_i^{\rm a}\Delta s_i^{\rm a} + \alpha^2 \Delta x_i \Delta s_i \geq \gamma(1-\alpha)\mu_g^k.$$

By using (6) one also has $\mu \leq \mathcal{O}\left((\mu_g^k)^4\right)$. The worst asymptotic value for α might be the result of the case when $x_i s_i = \gamma\mu_g$ and $\Delta x_i^{\rm a}\Delta s_i^{\rm a} > 0$. Assuming this, it is not clear whether $\Delta x_i \Delta s_i$ is nonnegative or negative. In case of nonnegativity, the previous inequality holds for a positive value of α. However, if $\Delta x_i \Delta s_i < 0$, it might not hold due to the very small μ value which is the result of sufficiently small μ_g. Therefore, modifying Mehrotra's heuristic might be cast as a way of achieving the superlinear convergence. The new adaptive updating strategy is defined by

$$\mu = \frac{\gamma t + \gamma \min(\mu_g^{\frac{1}{2}}, 1)}{1-\gamma}\mu_g, \tag{16}$$

where t is given by (11) and $0 < \gamma < \frac{1}{3}$. The "γt" term in this definition guarantees the existence of a positive step size following the proof of Theorems 3.3 and 4.4. However, the second term enables us to prove the superlinear convergence as it will be proven in what follows, since for small μ_g it is not as aggressive as (6). The "$1-\gamma$" term in the denominator of (16) is used for simplicity of the theoretical analysis which follows. Following the analysis of sections 3 and 4, changing Mehrotra's updating scheme to this updating strategy in Algorithms 1 and 2, while preserving the large update safeguard, does not change the order of the iteration complexity. Since the

large update safeguard gives us a positive lower bound for the maximum step size in the corrector step, for simplicity those complexity proofs are omitted here.

THEOREM 5.1. *Let the iterate (x^k, y^k, s^k) be generated by Algorithm* 1 *or* 2*, where μ is given by* (16)*. When μ_g is sufficiently small, Algorithms* 1 *and* 2 *are superlinearly convergent in the sense that $\mu_g^{k+1} = \mathcal{O}\left((\mu_g^k)^{1+r}\right)$ for some $r \in (0,1)$.*

Proof. Since $\left|(\Delta x^{\mathrm{a}})_i^k (\Delta s^{\mathrm{a}})_i^k\right| \leq \mathcal{O}((\mu_g^k)^2)$ for all $i \in \mathcal{I}$, then similar to the proof of Theorem 7.4 of [26] one can show that

$$|(\Delta x)_i^k (\Delta s)_i^k| \leq \mathcal{O}\left((\mu_g^k)^2\right).$$

By the new definition of μ, the next iterate is in the neighborhood $\mathcal{N}_\infty^-(\gamma)$ if for each $i \in \mathcal{I}$

$$(1-\alpha)x_i^k s_i^k + \alpha(1-\gamma)\mu^k - \alpha\Delta x_i^{\mathrm{a}}\Delta s_i^{\mathrm{a}} + \alpha^2\Delta x_i\Delta s_i \geq \gamma(1-\alpha)\mu_g^k. \tag{17}$$

Our goal is to find $\alpha \in (0,1]$ for which (17) holds. For this, it is sufficient to prove (17) for the case where $(\Delta x^{\mathrm{a}})_i^k (\Delta s^{\mathrm{a}})_i^k > 0$ and $(\Delta x)_i^k (\Delta s)_i^k < 0$. Using the definition of t, for a positive component of $\Delta x^a \Delta s^a$ one also has $\Delta x_i^{\mathrm{a}}\Delta s_i^{\mathrm{a}} \leq t x_i s_i$. Therefore, it suffices to find $\alpha \in (0,1]$ for which the following inequality holds:

$$(1-\alpha(1+t))x_i^k s_i^k + \alpha(1-\gamma)\mu^k - \alpha^2\mathcal{O}((\mu_g^k)^2) \geq \gamma(1-\alpha)\mu_g^k. \tag{18}$$

If (18) holds for $\alpha \geq \frac{1}{1+t}$, then $\alpha \geq 1-\mathcal{O}(\mu_g^k)$, since $\frac{1}{1+t} = \frac{1}{1+\mathcal{O}(\mu_g^k)} \geq 1-\mathcal{O}(\mu_g^k)$. Now let us assume that $\alpha < \frac{1}{1+t}$. In order to have (18), using the fact that $x_i^k s_i^k \geq \gamma\mu_g^k$, it suffices to have

$$\begin{aligned}(1-\alpha(1+t))\gamma\mu_g^k \;+\; \alpha\gamma t\mu_g^k + \alpha\gamma\min\left(\mu_g^{\frac{1}{2}}, 1\right)\mu_g - \alpha^2\mathcal{O}((\mu_g^k)^2) \\ \geq \gamma(1-\alpha)\mu_g^k\end{aligned}$$

for some $\alpha \in (0,1]$, which is equivalent to

$$\gamma\mu_g^{\frac{3}{2}} - \alpha\mathcal{O}((\mu_g^k)^2) \geq 0. \tag{19}$$

Inequality (19) definitely holds for $\alpha \geq 1-\mathcal{O}\left((\mu_g^k)^r\right)$, where $r \in (0,1)$.

Now, by using (10), one further has

$$\mu_g^k(\alpha_{\mathrm{c}}^k) = \left(1-\alpha_c^k(1-\mathcal{O}(\mu_g^k))\right)\mu_g^k \leq \left(1-(1-\mathcal{O}((\mu_g^k)^r))(1-\mathcal{O}(\mu_g^k))\right)\mu_g^k \leq \mathcal{O}\left((\mu_g^k)^{1+r}\right).$$

This gives the superlinear convergence of Algorithm 1 with the new choice of the parameter μ. The superlinear convergence of Algorithm 2 also can be proved analogously. ∎

6. Numerical results. In this section we report some illustrative numerical results for different variants of Algorithm 2 due to its better computational performance than Algorithm 1 for few problems. The results are obtained by modifying some of the subroutines of the McIPM (a self-dual embedding model-based implementation) and the LIPSOL (an infeasible IPM implementation for LO problems), two software packages based in MATLAB [29, 30]. Our computational experiments are done on a Pentium 4 machine with 2.53 GHZ and 512 MB ram. Numerical results are reported for all feasible NETLIB and Kennington test problems. For each problem we report the number of iterations, the time it takes to load the problem and solve it, and the number of exact digits, respectively.

For the McIPM package we use the following abbreviations for the different implementations of Mehrotra's algorithm:

- **PMMcIPM:** Mehrotra's original algorithm presented in section 1.
- **HMcIPM:** Mehrotra's original algorithm presented in section 1 combined with heuristics in the definition of the centering parameter. The interested reader can consult the McIPM package for heuristics that are used there [30].
- **NMcIPMI:** Algorithm 2.
- **NMcIPMII:** Algorithm 2 with the new definition of the centering parameter (16) instead of using (6).

For all the above-mentioned variants we set $\gamma = 10^{-4}$ and $\mu = \frac{\mu_g}{10}$ as the safeguard. In the implementation of our new definition of μ given by (16), we use $\mu = \frac{1}{5}(t + \min(\mu^1/2_g, 1))\mu_g$ rather than using (16), which is introduced for theoretical easiness.

Tables 1 to 3 show that for 36 problems (total number of problems is 112), the number of iterations for PMcIPM is higher than NMcIPMI, for 63 problems is higher than NMcIPMII, and for 65 problems is higher than HMcIPM implementations. As one can notice from Tables 1 to 3, for some problems PMcIPM is doing better than the other implementations, and for the rest they all perform equally. Significant difference in time occurs when the number of iterations is significantly different; for example, see "dfl001" and "degen3" in Table 1 and "osa-60" and "pds-20" in Table 3. The comparison between our two new algorithms (NMcIPMI and NMcIPMII) shows that NMcIPMII is better than NMcIPMI for 56 problems, while NMcIPMI is doing better only for 22 problems, and they perform equally on the rest of the problems. Therefore, overall NMcIPMII performs better than NMcIPMI. Finally, the comparison between NMcIPMII and HMcIPM shows that HMcIPM is doing better for 26 problems, while HMcIPM is better for 27 problems, and they perform equally on the rest of the problems. This comparison also shows that our simple safeguard-based algorithm is at least as effective as the heuristics used in the package and sometimes overperforms HMcIPM on difficult problems as given in Table 3.

The following abbreviations are also used for different implementations of Mehrotra's algorithm in LIPSOL:

- **PLIPSOL:** Infeasible variant of Mehrotra's algorithm presented in section 1.
- **HLIPSOL:** Infeasible variant of Mehrotra's original algorithm presented in section 1 with heuristics that are used in the definition of the centering parameter by LIPOSL. One should consult the LIPSOL package for the details of the heuristics [29].
- **SLIPSOL:** Infeasible variant of Algorithm 2.

For all the above-mentioned versions of LIPSOL we use $\gamma = 10^{-4}$ and $\frac{\beta}{1-\beta} = 10^{-1}$.

It is worthwhile mentioning that we do not have the second modification of the LIPSOL, namely the new definition of the centering parameter, because it requires a detailed analysis of the infeasible Mehrotra algorithm that is left for future research.

In Tables 4 to 6 we report the numerical results using the above-mentioned variants of LIPSOL. The comparison of iterations numbers show that for 66 problems SLIPSOL and HLIPSOL are doing better than PLIPSOL, while PLIPSOL is better only for a few problems. The comparison between SLIPSOL and HLIPSOL shows that overall they perform equally. Finally, a significant difference in time occurs when the number of iterations dramatically differ; for example, see "dfl001" in Table 4 and "cre-d," "osa-14," "pds-10," and "pds-20" in Table 6.

7. Final remarks. In this paper we have discussed the polynomiality of Mehrotra's original predictor-corrector algorithm. By a numerical example we have shown that Mehrotra's algorithm might lead to an inefficient algorithm while keeping the it-

TABLE 1
Comparison of the number of iterations for the NETLIB test problems.

Problem	PMMcIPM	NMcIPM I	NMcIPM II	HMcIPM	Problem	PMMcIPM	NMcIPM I	NMcIPM II	HMcIPM
25fv47	(28,4.17,7)	(28,4.07,7)	(28,4.5,7)	(27,4.42,6)	e226	(21,1,7)	(21,0.96,7)	(21,0.98,9)	(21,0.9,9)
80bau3b	(43,18.42,5)	(43,17.56,5)	(41,16.6,5)	(42,18.43,5)	etamacro	(26,1.9,8)	(25,1.7,8)	(24,1.5,7)	(25,1.7,7)
afiro	(10,0.34,8)	(10,0.27,8)	(10,0.31,8)	(10,0.3,9)	ffff800	(28,3,6)	(28,2.7,6)	(27,2.5,6)	(27,3,6)
adlittle	(14,0.33,8)	(14,0.37,8)	(14,0.33,8)	(14,0.42,9)	finnis	(30,2.1,5)	(27,2.1,5)	(28,2.1,5)	(28,2.1,5)
agg	(22,1.87,9)	(22,1.65,9)	(22,1.86,9)	(22,1.83,9)	fit1d	(26,3.13,6)	(24,2.67,6)	(24,3,8)	(26,2.94,8)
agg2	(19,2.31,8)	(19,1.72,8)	(19,1.76,8)	(19,2.11,11)	fit1p	(17,8.5,9)	(17,8.1,9)	(15,7.3,8)	(16,7.3,9)
agg3	(22,2.39,10)	(22,1.71,10)	(20,2.1,9)	(20,2.15,9)	fit2d	(24,23.6,9)	(22, 20.5,8)	(24,21.8,9)	(23,21.3,8)
bandm	(17,0.55,5)	(17,0.55,5)	(18,0.75,8)	(18,0.76,8)	fit2p	(21,23.5,10)	(20,21.8,8)	(21,22.9,8)	(21,23.8,8)
beaconfd	(12,0.61,7)	(12,0.5,7)	(12,0.51,7)	(12,0.71,7)	forplan	(30,1.8,4)	(29.1.1,5)	(30,1.53,4)	(31,1.8,5)
blend	(10,0.37,7)	(10,0.3,7)	(11,0.39,8)	(11,0.39,8)	ganges	(21,3.1,5)	(21,2.5,5)	(20,2.4,5)	(20,2.81,5)
bnl1	(33,2.71,7)	(33,2.6,7)	(33,2.5,7)	(32,2.35,7)	gfrd-pnc	(17,1.5,6)	(17,1.1,6)	(18,1.2,10)	(17,1.33,6)
bnl2	(41,11.86,7)	(41,11.86,7)	(38,10.72,7)	(37,10.91,7)	greenbea	(49,21,2)	(48,20,2)	(47,20.2,2)	(46,20,2)
boeing1	(25,1.94,7)	(25,1.98,7)	(24,1.8,7)	(24,1.94,7)	greenbeb	(48,17.15,4)	(47,17.5,4)	(46,16.2,4)	(48,17,4)
boeing2	(21,0.9,8)	(21,0.8,8)	(20,0.74,6)	(20,0.78,6)	grow15	(18,1.99,9)	(18,1.6,9)	(18,1.56,8)	(17,1.71,7)
bore3d	(21,0.84,9)	(20,0.82,9)	(18,0.74,9)	(17,0.72,7)	grow22	(19,2.5,8)	(19,2.38,8)	(19,2.6,8)	(18,2.35,8)
brandy	(18,1,7)	(18,0.8,7)	(17,0.6,7)	(17,0.63,7)	grow7	(18,1,8)	(18,1,8)	(18,1.2,7)	(17,1.1,7)
capri	(18,1.28,6)	(18,0.92,6)	(19,1.16,6)	(19,1.08,6)	israel	(22,1.9,8)	(22,1.74,6)	(22,1.6,7)	(22,1.88,6)
cycle	(40,12.23,6)	(38,11.72,6)	(39,11.63,6)	(39,11.77,6)	kb2	(18,0.57,10)	(18,0.38,10)	(18,0.49,10)	(17,0.35,8)
czprob	(32,3.71,9)	(32,3.4,8)	(32,3.7,8)	(32,3.75,11)	lotfi	(23,0.93,6)	(23,0.65,6)	(22,0.8,6)	(22,0.74,5)
d2q06c	(46,25.3,7)	(44,24.65,7)	(45,23.54,8)	(45,24.7,8)	maros-r7	(16,91,8)	(16,91,8)	(15,86,9)	(16,91,8)
d6cube	(20,6.35,9)	(20,6.73,10)	(20,6.5,10)	(20,6.1,10)	maros	(31,3.8,5)	(31,3.85,5)	(31,3.5,5)	(32,4.1,5)
degen2	(13,1.2,10)	(13,1.55,9)	(13,1.4,11)	(12,1.2,9)	modszk1	(29,2,4)	(29,2,4)	(29,2.1,6)	(29,2.2,6)
degen3	(43,23,8)	(22,13,8)	(14,9.5,8)	(14,9.5,8)	nesm	(34,6.3,6)	(34,6,6)	(32,5.94,6)	(31,5.8,6)
df1001	(49,607,6)	(47,584,,6)	(45,551,6)	(46,571,6)	perold	(48,5,5)	(48,5.6,6)	(42,4,6)	(42,5.1,5)

TABLE 2
Comparison of the number of iterations for the NETLIB test problems.

Problem	PMMcIPM	NMcIPM I	NMcIPM II	HMcIPM	Problem	PMMcIPM	NMcIPM I	NMcIPM II	HMcIPM
pilot	(54,44.52,4)	(65,52.55,4)	(49,39,4)	(48,39.51,4)	sctap3	(14,2.4,9)	(14,2.4,9)	(14,2.4,10)	(14,1.8,10)
pilot4	(39,4.61,6)	(37,3.98,6)	(35,3.46,6)	(37,4,6)	seba	(31,4.36,8)	(31,4.14,9)	(23,3.58,9)	(24,3.63,9)
pilotja	(49,10.25,5)	(44,9.83,5)	(42,8.75,6)	(43,9.25,6)	share1b	(28,0.7,5)	(28,0.7,5)	(27,0.83,5)	(27,0.8,7)
pilotnov	(28,5.9,9)	(28,6.15,10)	(26,5,9)	(26,5.57,9)	share2b	(12,0.45,7)	(11,0.42,7)	(11,0.53)	(11,0.33,6)
pilotwe	(46,5.63,6)	(44,5.29,6)	(42,5.4,6)	(42,5.33,6)	shell	(24,1.7,9)	(23,1.5,8)	(25,2,9)	(24,1.99,9)
pilot87	(77,176,5)	(71,161,5)	(71,158,5)	(79,178,5)	ship04l	(17,1.33,9)	(17,1.34,9)	(16,1.3,9)	(16,1.33,9)
recipe	(12,0.5,9)	(12,0.6,9)	(12,0.5,9)	(12,0.4,9)	ship04s	(17,1,7)	(18,1,7)	(17,0.9,7)	(16,1.1,7)
sc105	(12,0.4,6)	(12,0.4,6)	(11,0.31,6)	(12,0.28,6)	ship08l	(20,2.7,8)	(20,3.2,8)	(19,2.7,10)	(19,2.8,8)
sc205	(12,0.36,6)	(12,0.5,6)	(12,0.31,6)	(12,0.33,6)	ship08s	(19,1.3,8)	(19,1.7,10)	(17,1.34,8)	(17,1.35,8)
sc50a	(11,0.23,7)	(11,0.3,7)	(10,0.27,6)	(11.0.3,7)	ship12l	(28,4.9,8)	(27,3.7,9)	(27,3.86,9)	(27,4.3,8)
sc50b	(10.0.2,8)	(10,0.3,8)	(9,0.24,7)	(10,0.25,8)	ship12s	(23,1.86,9)	(23,1.5,9)	(21,1.82,7)	(21,1.8,7)
scagr25	(17,0.86,9)	(17,0.74,9)	(16,0.83,8)	(15,0.86,7)	sierra	(20,3.6,10)	(19,2.9,9)	(18,2.9,9)	(18,3.16,8)
scagr7	(14,0.43,7)	(14,0.36,7)	(13,0.39,7)	(13,0.44,7)	stair	(17,1.48,6)	(17,1,6)	(18,1.42,7)	(18,1.4,6)
scfxm1	(24,1.07,6)	(23,1.01,7)	(25,1.41,8)	(24,1.27,7)	standata	(17,0.98,10)	(17,0.96,10)	(18,0.86,9)	(18,1.22,9)
scfxm2	(27,1.96,8)	(27,1.91,8)	(26,1.96,8)	(25,2.07,7)	standmps	(20,1.36,10)	(19,1,9)	(20,0.86,10)	(19,1.6,10)
scfxm3	(27,2.85,7)	(26,2.4,7)	(26,2.4,7)	(26,2.75,7)	stocfor1	(14,0.49,8)	(14,0.38,8)	(15,0.56,8)	(14,0.48,7)
scorpion	(14,0.5,8)	(14,0.55,8)	(14,0.6,8)	(14,0.55,8)	stocfor2	(31,4.49,8)	(31,3.58,8)	(32,3.8,8)	(33,4.82,7)
scrs8	(26,1.8,5)	(25,1.83,5)	(24,1.48,5)	(24,1.4,5)	stocfo3	(50,48.7,5)	(50,48.3,5)	(49,47,5)	(51,50.4,5)
scsd1	(11,0.5,9)	(11,0.65,9)	(10,0.43,7)	(10,0.47,7)	truss	(21,5.5,8)	(21,5.94,8)	(20,4.93,8)	(20,5.27,8)
scsd6	(12,0.8,8)	(12,0.85,8)	(13,0.8,8)	(12,0.6,8)	tuff	(20,1.52,6)	(20,1.25,6)	(19,1.13,6)	(19,1.58,7)
scsd8	(10,1.3,10)	(10,1.4,10)	(11,1.3,10)	(10,1.1,10)	vtpbase	(17,0.8,9)	(20,0.7,9)	(17,0.58,8)	(17,0.6,8)
sctap1	(19,0.7,9)	(19,0.92,9)	(19,0.7,10)	(18,0.57,10)	woodlp	(16,5.4,9)	(15,5.3,9)	(15,4.95,9)	(15,5.5,10)
sctap2	(14,1.9,9)	(14,2,9)	(13,1.7,8)	(13,1.4,8)	woodw	(25,7.7,10)	(25,6.5,10)	(25,6.93,9)	(24,7.5,8)

TABLE 3
Comparison of the number of iterations for the Kennington test problems.

Problem	MMcIPM	NMcIPM I	NMcIPM II	HMcIPM
cre-a	(28,8.3,8)	(29,8.3,8)	(27,7.1,8)	(29,8.3,8)
cre-b	(37,205,8)	(36,200,8)	(34,188.7,8)	(34,190,8)
cre-c	(31,7.9,8)	(31,7.7,8)	(32,7.5,8)	(32,7.9,8)
cre-d	(35,177.5,8)	(34,173.20,8)	(33,169,8)	(32,163.4,8)
ken-07	(17,4,8)	(17,3.2,8)	(17,4,7)	(17,3.6,7)
ken-11	(21,29.7,7)	(21,29.5,7)	(20,29.6,7)	(20,30,7)
ken-13	(29,91,7)	(28,88,7)	(27,88.5,7)	(26,83,7)
ken-18	(37,637.6,8)	(36,621.6,8)	(34,590.6,8)	(35,621.6,8)
osa-07	(31,31,8)	(31,31,8)	(34,33,8)	(40,39,7)
osa-14	(39,91,8)	(39,87.5,8)	(45,95.2,7)	(52,114,8)
osa-30	(44,,209.5,7)	(41,197.3,7)	(44,206,7)	(44,207,7)
osa-60	(51,612,7)	(48,580,8)	(47,573,8)	(57,668,8)
pds-02	(34,12,7)	(33,11,8)	(32,10.8,88)	(32,11,7)
pds-06	(51,164,7)	(50,160,7)	(43,136.5,7)	(45,146,7)
pds-10	(70,864,7)	(69,854,8)	(56,698,8)	(58,725,8)
pds-20	(96,8164.5,7)	(85,7552.6,7)	(79,6763,7)	(81,6926.4,7)

erate in the $\mathcal{N}_\infty^-(\gamma)$ neighborhood, which is essential to prove the polynomial iteration complexity. This motivated us to combine his idea with a safeguard strategy that allows us to get a positive lower bound for the step size in the corrector step. Further, by slightly changing the Newton system, the iteration complexity of the algorithm is significantly reduced. This also led us to superior computational performance of the algorithm. To ensure the superlinear convergence of the algorithm we changed Mehrotra's updating scheme of the centering parameter so that the new algorithms preserve the iteration complexity and exhibit stronger asymptotic convergence properties. Our illustrative numerical results show that our new safeguard-based algorithms are competitive with two state of the art software packages that are using heuristics to stabilize the convergence of the implemented algorithms.

There are several interesting questions regarding the proposed safeguard strategy. For example, one can analyze the infeasible variant of this prototype. It is also possible to extend this approach to other classes of optimization problems such as SDO, SOCO, and convex nonlinear optimization.

Appendix. In this section we prove two technical lemmas that have been used frequently during the analysis.

LEMMA A.1. *Let* $(\Delta x^a, \Delta y^a, \Delta s^a)$ *be the solution of* (4). *Then*

$$\Delta x_i^a \Delta s_i^a \leq \frac{x_i s_i}{4} \quad \forall i \in \mathcal{I}_+.$$

Proof. By (4) for $i \in \mathcal{I}_+$ we have

$$s_i \Delta x_i^a + x_i \Delta s_i^a = -x_i s_i.$$

TABLE 4
Comparison of the number of iterations for the Kennington test problems.

Problem	PLIPSOL	SLIPSOL	HLIPSOL	Problem	PLIPSOL	SLIPSOL	HLIPSOL
25fv47	(27,3.1,10)	(24,4,11)	(25,3.7,10)	e226	(23,1.2,7)	(20,0.98,7)	(21,1.2,11)
80bau3b	(46,12.1,4)	(40,10.8,4)	(39,11.3,4)	etamacro	(27,1.8,7)	(25,1.88,7)	(25,1.6,7)
afiro	(8,0.34,10)	(8,0.35,10)	(8,0.2,10)	ffff800	(31,2.95,6)	(26,2.5,6)	(26,2.75,6)
adlittle	(13,0.57,11)	(13,0.57,11)	(13,0.36,11)	finnis	(36,1.8,5)	(28,1.8,5)	(30,1.5,5)
agg	(21,1.4,10)	(21,1.64,10)	(21,1.45,10)	fit1d	(20,1.63,11)	(18,1.74,11)	(19,1.7,11)
agg2	(20,1.9,10)	(19,1.82,10)	(18,2.1,10)	fit1p	(17,12.6,10)	(16,12.3,10)	(16,11.7,10)
agg3	(19,1.86,11)	(18,1.77,11)	(17,1.82,11)	fit2d	(23,11.7,11)	(22,12,11)	(22,11.5,9)
bandm	(18,0.86,9)	(18,0.85,8)	(18,1,10)	fit2p	(21,28.5,9)	(21,30,9),	(21,33,9)
beaconfd	(13,0.7,11)	(13,0.68,11)	(13,0.78,11)	forplan	(25,1.25,6)	(24,1.1,6)	(22,1.2,6)
blend	(12,0.44,10)	(12,0.42,10)	(12,0.5,10)	ganges	(19,2,5)	(18,2,5)	(18,2,5)
bnl1	(33,2.14,7)	(30,2,7)	(26,2,7)	gfrd-pnc	(21,1,11)	(20,0.7,11)	(21,1,11)
bnl2	(38,12.3,9)	(31,10.1,9)	(31,10.4,9)	greenbea	(43,18,3)	(48,19.9,3)	4(43,18.8,2)
boeing1	(24,1.6,10)	(22,1.4,10)	(21,1.6,9)	greenbeb	(42,14.4,4)	(37,12.6,4)	(38,13.6,4)
boeing2	(20,0.83,10)	(17,0.65,8)	(19,1.17,8)	grow15	(18,1.44,11)	(17,1.3,11)	(17,1.4,11)
bore3d	(17,0.83,11)	(18,0.8,10)	(18,1,11)	groww22	(19,2,11)	(18,1.8,11)	(19,1.95,11)
brandy	(18,0.88,10)	(15,0.75,10)	(17,1.2,10)	grow7	(17,0.88,10)	(16,0.86,10)	(16,0.9,10)
capri	(18,1,9)	(18,1,10)	(20,1.2,10)	israel	(25,1.76,11)	(22,1.6,11)	(23,1.3,11)
cycle	(26,8.9,0)	(25,8.4,6)	(24,8.6,6)	kb2	(14,0.45,10)	(14,0.4,10)	(15,0.55,11)
czprob	(37,2.9,11)	(38,3,11)	(36,3,11)	lotfi	(18,0.9,7)	(18,1.1,10)	(18,0.8,10)
d2q06c	(35,25.5,7)	(33,23.9,7)	(32,24,7)	maros-r7	(16,154,8)	(15,151,11)	(15,144,11)
d6cube	(27,8.1,9)	(24,7.5,7)	(23,7.2,9)	maros	(31,7,5)	(33,4.4,11)	(33,4.2,11)
degen3	(26,20.4,9)	(20,16.9,11)	(20,16.4,8)	modszk1	(24,2,9)	(24,1.8,9)	(24,1.8,9)
degen2	(14,1.48,9)	(14,1.46,9)	(14,1.5,9)	nesm	(35,4.9,6)	(33,5,6)	(33,4.7,6)
df1001	(81,1883.33,6)	(59,1452.6,6)	(73,1810,6)	perold	(38,3.85,5)	(33,3.46,5)	(31,2.7,5)

If we divide this equation by $x_i s_i$ we get

$$\frac{\Delta x_i^a}{x_i} + \frac{\Delta s_i^a}{s_i} = -1.$$

Since $\Delta x_i^a \Delta s_i^a > 0$, this equality implies that both $\Delta x_i^a < 0$ and $\Delta s_i^a < 0$. Then from

$$0 \le \left(\frac{\Delta x_i^a}{x_i} - \frac{\Delta s_i^a}{s_i}\right)^2 = \left(\frac{\Delta x_i^a}{x_i}\right)^2 + \left(\frac{\Delta s_i^a}{s_i}\right)^2 - 2\frac{\Delta x_i^a \Delta s_i^a}{x_i s_i} = 1 - 4\frac{\Delta x_i^a \Delta s_i^a}{x_i s_i}$$

we get

$$\Delta x_i^a \Delta s_i^a \le \frac{x_i s_i}{4}. \qquad \square$$

LEMMA A.2. *Let* $(\Delta x^a, \Delta y^a, \Delta s^a)$ *be the solution of* (4)*; then we have*

$$\sum_{i \in \mathcal{I}_+} \Delta x_i^a \Delta s_i^a = \sum_{i \in \mathcal{I}_-} |\Delta x_i^a \Delta s_i^a| \le \frac{1}{4} \sum_{i \in \mathcal{I}_+} x_i s_i \le \frac{x^T s}{4}.$$

TABLE 5
Comparison of the number of iterations for the NETLIB test problems.

Problem	PLIPSOL	SLIPSOL	HLIPSOL	Problem	PLIPSOL	SLIPSOL	HLIPSOL
pilot	(39,43.6,4)	(30,33.6,4)	(31,37,4)	sctap3	(19,1.92,11)	(18,2.3,12)	(18,2.3,12)
pilot4	(31,2.95,8)	(34,3.2,8)	(30,2.65,8)	seba	(20,3.64,12)	(22,4.8,12)	(22,4.3,12)
pilotja	(34,7.4,6)	(32,7,6)	(31,6.7,6)	share1b	(23,0.7,11)	(22,0.83,11)	(22,0.88,11)
pilotnov	(21,4.1,11)	(19,3.94,11)	(20,4.4,11)	share2b	(13,0.47,11)	(12,0.4,11)	(13,0.55,11)
pilotwe	(40,3.5,6)	(35,3,6)	(37,3.8,6)	shell	(23,1.1,11)	(19,0.78,11)	(21,1.27,11)
pilot87	(42,159,6)	(38,151,6)	(38,149,6)	ship04l	(14,0.97,9)	(14,0.9,9)	(14,1,9)
recipe	(9,0.45,11)	(9,0.53,11)	(9,0.58,11)	ship04s	(15,0.86,11)	(14,0.75,11)	(14,0.5,11)
sc105	(10,0.5,11)	(10,0.4,11)	(10,0.4,11)	ship08l	(16,1.9,11)	(16,1.8,11)	(16,1.8,11)
sc205	(11,0.5,11)	(11,0.4,11)	(11,0.4,11)	ship08s	(15,1.14,11)	(15,1.1,11)	(15,1,11)
sc50a	(10,0.33,10)	(10,0.45,10)	(10,0.4,10)	ship12l	(17,1.9,1)	(19,2,11)	(18,1.6,11)
sc50b	(7,0.38,8)	(7,0.33,8)	(7,0.27,8)	ship12s	(18,1,11)	(18,1,11)	(18,0.8,11)
scagr25	(17,0.8,)	(17,0.7,)	(17,0.95,)	sierra	(18,1.95,10)	(18,1.72,10)	(17,2,10)
scagr7	(14,0.52,7)	(13,0.44,7)	(14,0.63,7)	stair	(16,1.4,11)	(14,1.15,10)	(14,1.4,10)
scfxm1	(20,0.92,11)	(19,0.78,10)	(19,1.1,11)	standata	(18,0.92,11)	(16,0.75,11)	(17,1,11)
scfxm2	(22,1.55,11)	(20,1.4,11)	(21,1.7,10)	standmps	(25,1.25,11)	(23,1.2,9)	(24,1.46,11)
scfxm3	(23,2,10)	(20,1.72,10)	(21,2.1,10)	stocfor1	(15,0.53,11)	(17,0.42,11)	(16,0.67,11)
scorpion	(15,0.64,11)	(15,0.65,11)	(15,0.6,11)	stocfor2	(22,2.75,10)	(23,2.85,10)	(21,2.94,10)
scrs8	(25,1.4,5)	(25,1.55,5)	(24,1.35,5)	stocfor3	(33,30.3,5)	(33,30.3,5)	(33,30.3,5)
scsd1	(10,0.52,11)	(10,0.63,7)	(9,0.7,5)	truss	(20,4.1,10)	(18,3.8,10)	(19,4.1,10)
scsd6	(12,0.66,10)	(11,0.78,6)	(12,0.8,10)	tuff	(23,1.44,6)	(20,0.95,6)	(20,1.5,4)
scsd8	(11,0.91,10)	(11,1.1,10)	(11,1.16,10)	vtpbase	(19,0.72,11)	(27,0.81,11)	(23,1,11)
sctap1	(19,0.7,12)	(17,0.8,12)	(17,0.7,12)	woodlp	(24,6.15,10)	(19,5.3,10)	(19,5.3,6)
sctap2	(17,1.55,11)	(16,1.8,10)	(19,1.6,10)	woodw	(30,7.3,5)	(26,6,6)	(28,7.1,8)

TABLE 6
Comparison of Iterations Number for the Kennington Test Problems.

Problem	PLIPSOL	SLIPSOL	HLIPSOL
cre-a	(33,7.18,8)	(29,6.3,8)	(30,6.8,8)
cre-b	(45,352,8)	(37,304.4)	(42,332.4,8)
cre-c	(32,6,8)	(31,6.1,8)	(30,6,8)
cre-d	(47,310.5,8)	(38,264.6,8)	(38,271,8)
ken-07	(16,2.2,8)	(15,2.2,8)	(16,2.2,8)
ken-11	(23,18.8,7)	(21,18.1,7)	(22,18.3,7)
ken-13	(30,60.7,10)	(28,60,10)	(27,55.3,10)
ken-18	(42,650,8)	(42,632,8)	(38,574,8)
osa-07	(29,25.5,7)	(27,25.7,7)	(27,24.4,7)
osa-14	(30,63.5,8)	(34,74,8)	(37,76,8)
osa-30	(37,160.2,8)	(39,183,8)	(36,157,8)
osa-60	(39,461,8)	(38,451,8)	(34,422,8)
pds-02	(29,6.75,7)	(28,7.2,7)	(29,7.1,7)
pds-06	(44,200,7)	(45,211,7)	(42,191,7)
pds-10	(58,1226,8)	(55,1183,8)	(52,1104,8)
pds-20	(69,10975.2,7)	(63,10028,7)	(67,10645,7)

Proof. Since $(\Delta x^a)^T \Delta s^a = 0$, the proof is a direct consequence of Lemma A.1. □

Acknowledgments. The authors thank the associate editor and two anonymous referees for their valuable comments on the earlier versions of this paper.

REFERENCES

[1] E. D. ANDERSEN AND K. D. ANDERSEN, *The MOSEK interior point optimizer for linear programming: An implementation of the homogeneous algorithm*, in High Performance Optimization, H. Frenk, K. Roos, T. Terlaky, and S. Zhang, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000, pp. 197–232.

[2] K. M. ANSTREICHER AND R. A. BOSCH, *A new infinity-norm path following algorithm for linear programming*, SIAM J. Optim., 5 (1995), pp. 236–246.

[3] C. CARTIS, *Some Disadvantages of a Mehrotra-type Primal-Dual Corrector Interior Point Algorithm for Linear Programming*, 2005, http://www.optimization-online.org.

[4] M. COLOMBO AND J. GONDZIO, *Further Development of Multiple Centrality Correctors for Interior Point Methods*, 2005, http://www.optimization-online.org.

[5] J. CZYZYK, S. MEHRTOTRA, M. WAGNER, AND S. J. WRIGHT, *PCx: An interior-point code for linear programming*, Optim. Methods Softw., 11/12 (1999), pp. 397–430.

[6] J. GONDZIO, *Multiple centrality corrections in a primal-dual method for linear programming*, Comput. Optim. Appl., 6 (1996), pp. 137–156.

[7] C. C. GONZAGA, *Complexity of predictor-corrector algorithms for LCP based on a large neighborhood of the central path*, SIAM J. Optim., 10 (1999), pp. 183–194.

[8] P.-F. HUNG AND Y. YE, *An asymptotical $O(\sqrt{n}L)$-iteration path-following linear programming algorithm that uses wide neighborhoods*, SIAM J. Optim., 6 (1996), pp. 570–586.

[9] F. JARRE AND M. WECHS, *Extending Mehrotra's corrector for linear programs*, Adv. Model. Optim., 1 (1999), pp. 38–60.

[10] J. JI, F. A. POTRA, AND S. HUANG, *A predictor-corrector method for linear complementarity problems with polynomial complexity and superlinear convergence*, J. Optim. Theory Appl.,

84 (1995), pp. 187–199.
[11] N. K. Karmarkar, *A new polynomial-time algorithm for linear programming*, Combinatorica, 4 (1984), pp. 373–395.
[12] N. Megiddo, *Pathways to the optimal set in linear programming*, in Progress in Mathematical Programming: Interior Point and Related Methods, N. Megiddo, ed., Springer-Verlag, New York, 1989, pp. 131–158.
[13] S. Mehrotra, *On finding a vertex solution using interior-point methods*, Linear Algebra Appl., 152 (1991), pp. 233–253.
[14] S. Mehrotra, *On the implementation of a primal-dual interior point method*, SIAM J. Optim., 2 (1992), pp. 575–601.
[15] S. Mehrotra and Z. Li, *Convergence conditions and Krylov subspace–based corrections for primal-dual interior-point method*, SIAM J. Optim., 15 (2005), pp. 635–653.
[16] S. Mizuno, M. J. Todd, and Y. Ye, *On adaptive step primal-dual interior-point algorithms for linear programming*, Math. Oper. Res., 18 (1993), pp. 964–981.
[17] R. D. C. Monteiro, I. Adler, and M. G. C. Resende, *A polynomial-time primal-dual affine scaling algorithm for linear and convex quadratic programming and its power series extensions*, Math. Oper. Res., 15 (1990), pp. 191–214.
[18] J. Peng, T. Terlaky, and Y. Zhao, *A predictor-corrector algorithm for linear optimization based on a specific self-regular proximity function*, SIAM J. Optim., 15 (2005), pp. 1105–1127.
[19] F. A. Potra, *The Mizuno-Todd-Ye algorithm in a large neighborhood of the central path*, European J. Oper. Res., 143 (2002), pp. 257–267.
[20] F. A. Potra, *A superlinearly convergent predictor-corrector method for degenerate LCP in a wide neighborhood of the central path*, Math. Program., 100 (2004), pp. 317–337.
[21] F. A. Potra and X. Liu, *Predictor-corrector methods for sufficient linear complementarity problems in a wide neighborhood of the central path*, Optim. Methods Softw., 20 (2005), pp. 145–168.
[22] C. Roos, T. Terlaky, and J.-Ph.Vial, *Theory and Algorithms for Linear Optimization: An Interior Point Approach*, John Wiley and Sons, Chichester, UK, 1997.
[23] M. Salahi and T. Terlaky, *Adaptive large neighborhood self-regular predictor-corrector IPMs for LO*, J. Optim. Theory Appl., 132 (2007), pp. 143–160.
[24] G. Sonnevend, *An "analytic center" for polyhedrons and new classes of global algorithms for linear (smooth, convex) programming*, in System Modeling and Optimization: Proceedings of the 12th IFIP Conference (Budapest 1985), Lecture Notes in Control Inform. Sci. 84, A. Prékopa, J. Szelezsán, and B. Strazicky, eds., Springer-Verlag, Berlin, 1986, pp. 866–876.
[25] R. Tapia, Y. Zhang, M. Saltzman, and A. Weiser, *The Mehrotra predictor-corrector interior-point method as a perturbed composite Newton method*, SIAM J. Optim., 6 (1996), pp. 47–56.
[26] S. J. Wright, *Primal-Dual Interior-Point Methods*, SIAM, Philadelphia, 1997.
[27] Y. Ye, *Interior Point Algorithms, Theory and Analysis*, John Wiley and Sons, Chichester, UK, 1997.
[28] Y. Zhang and D. Zhang, *On the polynomiality of the Mehrotra-type predictor-corrector interior point algorithms*, Math. Programming, 68 (1995), pp. 303–317.
[29] Y. Zhang, *Solving large-scale linear programs by interior point methods under the MATLAB enviroment*, Optim. Methods Softw., 10 (1999), pp. 1–31.
[30] X. Zhu, J. Peng, T. Terlaky, and G. Zhang, *On Implementing Self-Regular Proximity Based Feasible IPMs*, Technical report 2003/2, Advanced Optimization Lab, Department of Computing and Software, McMaster University, Hamilton, ON, Canada, http://www.cas.mcmaster.ca/∼oplab/publication.

SIAM J. OPTIM.
Vol. 18, No. 4, pp. 1398–1413

EFFICIENT REDUCTION OF POLYNOMIAL ZERO-ONE OPTIMIZATION TO THE QUADRATIC CASE*

CHRISTOPH BUCHHEIM† AND GIOVANNI RINALDI‡

Abstract. We address the problem of optimizing a polynomial with real coefficients over binary variables. We show that a complete polyhedral description of the linearization of such a problem can be derived in a simple way from the polyhedral description of the linearization of some quadratic optimization problem. The number of variables in the latter linearization is only slightly larger than in the former. If polynomial constraints are present in the original problem, then their linearized counterparts carry over to the linearized quadratic problem unchanged. If the original problem formulation does not contain any constraints, we obtain a reduction to unconstrained quadratic zero-one optimization, which is equivalent to the well-studied max-cut problem. The separation problem for general unconstrained polynomial zero-one optimization thus reduces to the separation problem for the cut polytope. This allows us to transfer the entire knowledge gained for the latter polytope by intensive research and, in particular, the sophisticated separation techniques that have been developed. We report preliminary experimental results obtained with a straightforward implementation of this approach.

Key words. polynomial zero-one optimization, integer nonlinear programming, pseudo-boolean functions, max-cut problem, multilinear function optimization

AMS subject classifications. 90C57, 65K05

DOI. 10.1137/050646500

1. Introduction.

We consider the problem of maximizing (or minimizing) a polynomial objective function over binary variables under arbitrary polynomial constraints, i.e., over all binary vectors belonging to a given semialgebraic set. We call it *polynomial zero-one optimization*.

Many different names and interpretations of this problem are circulating. For instance, if a set A is used to index the (original) variables, then one might think of the variables as modeling some subset of A. The possible monomials are then in one-to-one correspondence with the subsets of A, as we may assume that every variable appears with exponent at most one in every monomial. Consequently, maximizing a polynomial over these variables means searching for a subset S of A such that the sum of all coefficients of monomials that correspond to subsets of S is maximized. In other words, for every subset of A, one can arbitrarily reward or punish the fact that all elements of this set are chosen by S.

Still in the unconstrained case, assume that the objective function is multilinear. Then the constraints $x \in \{0,1\}^A$ can be replaced by $x \in [0,1]^A$, as an optimal solution is always attained at a vertex of the unit hypercube [25]. Moreover, as multilinear functions are closed under affine variable transformations, the unit hypercube can be replaced by an arbitrary box. Therefore, this case is equivalent to the problem of *maximizing a multilinear function over a box*.

*Received by the editors December 1, 2005; accepted for publication (in revised form) July 7, 2007; published electronically December 21, 2007. This work was partially supported by the Marie Curie Research Training Network 504438 (ADONET) funded by the European Commission.
http://www.siam.org/journals/siopt/18-4/64650.html
†Institut für Informatik, Universität zu Köln, Pohligstr. 1, 50969 Köln, Germany (buchheim@informatik.uni-koeln.de).
‡Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti" del CNR, viale Manzoni 30, 00185 Roma, Italy (rinaldi@iasi.cnr.it).

A standard method to address such a problem is linearization: every nonlinear term in the problem formulation is replaced by a newly introduced binary variable. In other words, the new variables correspond to all monomials appearing in the objective function or in one of the constraints of the original problem. Over this new set of variables, the problem thus translates to an integer linear program, which can hopefully be solved by polyhedral methods.

However, additional constraints are now needed to link the newly introduced variables to the original ones. More precisely, the value of a variable representing a monomial $x_{i_1}x_{i_2}\dots x_{i_r}$ must equal the product of the values of $x_{i_1}, x_{i_2}, \dots, x_{i_r}$. The purpose of this paper is to show that, after adding a small number of new variables, this task can be reduced to the quadratic case in an efficient (and easily implementable) way. Using elementary polyhedral results, we show that the separation problem for the linearized original problem can be reduced to the separation problem for a linearized quadratic problem that is only slightly larger. The construction of the latter is "orthogonal" to the given polynomial constraints in the sense that their linearized counterparts carry over to the quadratic problem unchanged. By the equivalence of unconstrained quadratic zero-one optimization and max-cut, we can thus reduce the original problem to a max-cut problem with additional linear constraints.

The most promising application for this reduction method arises in the unconstrained case, which we examine in more detail in this paper, also experimentally. In this case, the separation problem reduces to the separation problem for the cut polytope, without further constraints. The cut polytope is one of the most important and best-studied objects in polyhedral combinatorics: many classes of cutting planes as well as sophisticated separation techniques are at hand for addressing the corresponding separation problem (see, e.g., [19] or [21]). Our approach aims at exploiting this knowledge for optimizing polynomials of arbitrary degree.

The polynomial zero-one optimization problem, with different types of constraints, has been studied extensively in the literature, often under the name of *pseudo-boolean optimization*. The reader is referred to a comprehensive survey by Boros and Hammer [5] that contains not only pointers to numerous applications but also approaches for solving this problem. In particular, two different approaches are mentioned: a reduction to the quadratic case due to Rosenberg [26] and the so-called *basic algorithm* by Hammer, Rosenberg, and Rudeanu [16]. The former approach can be combined with any max-cut algorithm and is briefly discussed in section 3.5. For the latter approach, to the best of our knowledge, an experimental evaluation exists only for a special case where it runs in linear time [7]. The ongoing interest in pseudo-boolean optimization is underlined by a recent special issue of the journal *Discrete Applied Mathematics* [14]. We would also like to point to the numerous theoretical results that have been obtained for pseudo-boolean functions, e.g., concerning persistency [15], and to interesting special cases such as hyperbolic pseudo-boolean functions or products of linear functions; see again [5].

This paper is organized as follows. In section 2, we define the general problem addressed and recall the standard linearization method. In section 3, we present the reduction method for polynomial zero-one programming to the quadratic case. The unconstrained case is considered in section 4; in particular, preliminary experimental results are reported. Some conclusions are drawn in section 5.

2. The standard linearization. Let A be any finite set of n elements, and consider the set of binary variables $\{x_a \mid a \in A\}$. Let $g_0, \dots, g_r$ be polynomials over

these variables. We consider the following polynomial zero-one optimization problem:

$$\begin{array}{rrcll} \max & g_0(x) & & & \\ \text{s.t.} & g_i(x) & \geq & 0 & \text{for all } i = 1, \dots, r, \\ & x & \in & \{0,1\}^A \, . & \end{array} \tag{2.1}$$

The standard way of linearizing the problem (2.1) is to consider the set $\mathcal{I}$ of all nonconstant monomials appearing in at least one of the g_i's and to use a binary variable z_I for every $I \in \mathcal{I}$, yielding the equivalent formulation

$$\begin{array}{rrcll} \max & h_0(z) & & & \\ \text{s.t.} & h_i(z) & \geq & 0 & \text{for all } i = 1, \dots, r, \\ & z_I & = & \prod_{a \in I} z_{\{a\}} & \text{for all } I \in \mathcal{I}, \\ & z & \in & \{0,1\}^{\mathcal{I}} \, . & \end{array} \tag{2.2}$$

Here we use h_i to denote the polynomial g_i considered as a linear function in the new set of variables z_I. As all variables in (2.1) are binary, we may assume that each g_i is multilinear. In particular, we can identify monomials with the corresponding subsets of A and thus consider $\mathcal{I}$ as a subset of $2^A \setminus \{\emptyset\}$; therefore, we will use expressions such as "union of monomials" in the following. Moreover, for the ease of exposition, we assume that $\{a\} \in \mathcal{I}$ for all $a \in A$. We denote the cardinality of $\mathcal{I}$ by m.

Throughout this paper, let F denote the set of feasible solutions of (2.2), and let P denote the convex hull of F, which is a polytope in $\mathbb{R}^{\mathcal{I}}$. Problem (2.2) is still a polynomial optimization problem, but the nonlinearity is restricted to the product formulae. The remaining problem is to model these constraints by linear inequalities. The easiest way to do this is the following: for every $I \in \mathcal{I}$ with $|I| \geq 2$, use the $|I|+1$ linear inequalities

$$z_I \leq z_{\{a\}} \quad \text{for all } a \in I \, , \tag{2.3}$$

$$z_I \geq \sum_{a \in I} z_{\{a\}} - |I| + 1 \, . \tag{2.4}$$

This standard linearization approach has been widely discussed in the literature; for early examples, see [10, 11, 12, 16, 28]. However, the relaxation of P given by the constraints (2.3) and (2.4) is rather weak when the integrality constraints are dropped. The aim of this paper is to present a general method for deriving much tighter relaxations of P.

A different approach to problem (2.1) is to replace each polynomial function g_i by a collection of linear functions in the same set of variables. This approach has been studied by Granot and Hammer [13] and later improved by Balas and Mazzola [2]. Unfortunately, the resulting LP has exponential size, and the corresponding relaxation is rather weak in general. On the other hand, our aim is to keep the construction as small as possible, at the same time allowing a tight polyhedral description.

Still another approach to solving problem (2.1) is to apply lift-and-project methods; for this topic, see the comparative survey by Laurent [20]. Here new polynomial constraints are created by multiplication of given ones. In the LP-based approaches, the resulting constraints are then linearized. According to the number of constraints multiplied, one obtains hierarchies of LP relaxations that coincide with the polytope P

at level n. Similar ideas have been used to develop cutting plane algorithms, e.g., by Balas et al., who proposed strengthening the standard linearization by applying the lifting to single fractional variables and computing most violated facets of the corresponding relaxations, always projecting them back to the original variable space in order to avoid exponential numbers of variables [1, 3].

Related methods based on semidefinite programming (SDP) have been introduced by Lasserre [18] and Parillo [24]. These are based on deep results from real algebraic geometry on representations of nonnegative polynomial as sums of squares. As shown in [20], the hierarchy of relaxations constructed by Lasserre is stronger than earlier hierarchies of LP-based relaxations [1, 22, 27]. However, in all these approaches the number of variables already becomes large in the first levels of the hierarchy.

3. The general reduction method. In the following, we present a general method of reducing polynomial zero-one optimization to the quadratic case via the separation problem. We first prove elementary polyhedral results that provide the theoretical background for the reduction (section 3.1). Then we give an algorithmic description of the reduction procedure (section 3.2). Special aspects of this method are discussed in sections 3.3 and 3.4. Finally, we compare this approach to the direct reduction method (section 3.5).

3.1. Basic results. Starting from the polytope P corresponding to (2.2), we first construct a new polytope P^* as follows: define the following subset of $2^{\mathcal{I}}$ whose elements are all sets of two (not necessarily distinct) monomials whose union is a monomial of $\mathcal{I}$:

$$\mathcal{I}^* = \left\{ \{I, J\} \mid I, J \in \mathcal{I} \text{ and } I \cup J \in \mathcal{I} \right\}.$$

As in the definition of $\mathcal{I}^*$ the sets I and J may coincide, we have $\{I\} \in \mathcal{I}^*$ for each $I \in \mathcal{I}$. If we associate a variable $y_{\{I,J\}}$ with every element $\{I, J\}$ of $\mathcal{I}^*$, every h_i gives rise to a linear function h_i^* over the new variables $\{y_S \mid S \in \mathcal{I}^*\}$ by replacing each variable z_I with the corresponding variable $y_{\{I\}}$. Now consider

$$\begin{array}{lrcll} \max & h_0^*(y) & & & \\ \text{s.t.} & h_i^*(y) & \geq & 0 & \text{for all } i = 1, \ldots, r, \\ & y_{\{I,J\}} & = & y_{\{I\}} y_{\{J\}} & \text{for all } \{I, J\} \in \mathcal{I}^*, \\ & y & \in & \{0,1\}^{\mathcal{I}^*}, & \end{array} \tag{3.1}$$

and let F^* be the set of feasible solutions of (3.1). Moreover, let P^* be the polytope in $\mathbb{R}^{\mathcal{I}^*}$ defined as the convex hull of F^*.

We do not aim to solve this optimization problem but rather are interested only in the separation problem for P^*. Note that the variables that appear in the objective function and in the r linear constraints are only those indexed by a single monomial. The variables indexed by sets of two monomials appear only in the product constraints. Thus problem (3.1) can be thought of as the linearization of a quadratic zero-one optimization problem. Our aim is to exploit this in the following. To do so, we define an injective linear map $\psi: \mathbb{R}^{\mathcal{I}} \to \mathbb{R}^{\mathcal{I}^*}$ componentwise by $(\psi(z))_{\{I,J\}} = z_{I \cup J}$, and we set

$$Y = \left\{ y \in \mathbb{R}^{\mathcal{I}^*} \mid y_{\{I \cup J\}} = y_{\{I,J\}} \text{ for all } I, J \in \mathcal{I} \text{ with } I \cup J \in \mathcal{I} \right\}.$$

From these definitions, it is readily checked that $Y = \psi(\mathbb{R}^{\mathcal{I}})$. The preimage z under ψ of a vector y in Y is determined by $z_I = y_{\{I\}}$.

LEMMA 3.1. *$\psi(F) \subseteq F^* \cap Y$.*

Proof. Let $z \in F$, and set $y = \psi(z)$. By definition, we have $h_i^*(y) = h_i(z)$, so that all inequalities in (3.1) are satisfied by y. It remains to show $y_{\{I,J\}} = y_{\{I\}} y_{\{J\}}$ for all $I, J \in \mathcal{I}$. Indeed, as z belongs to F, we have

$$y_{\{I,J\}} = z_{I \cup J} = \prod_{a \in I \cup J} z_{\{a\}} = \prod_{a \in I} z_{\{a\}} \prod_{a \in J} z_{\{a\}} = z_I z_J = y_{\{I\}} y_{\{J\}} \,. \qquad \square$$

Lemma 3.1 shows that every LP relaxation of P^* gives rise to an LP relaxation of P. The following results concern the quality of this relaxation. For this, call the set $\mathcal{I}$ *reducible* if every set in $\mathcal{I}$ containing more than one element is a union of two other sets in $\mathcal{I}$. Then we have the following.

THEOREM 3.2. *If $\mathcal{I}$ is reducible, then $\psi(F) = F^* \cap Y$.*

Proof. By Lemma 3.1, it remains to show that every $y \in F^* \cap Y$ belongs to $\psi(F)$. As mentioned above, the vector z defined by $z_I = y_{\{I\}}$ satisfies $\psi(z) = y$, and so we have to show that z belongs to F. As $h_i(z) = h_i^*(y)$, all inequalities in (2.2) are satisfied by z. As z is integer, it remains to show that

$$z_I = \prod_{a \in I} z_{\{a\}}$$

holds for all $I \in \mathcal{I}$. By definition of z, we thus have to prove

$$y_{\{I\}} = \prod_{a \in I} y_{\{\{a\}\}}$$

for all $I \in \mathcal{I}$. This is done by induction over the cardinality of I. If $|I| = 1$, the claim is trivial. Otherwise, as $\mathcal{I}$ is reducible, there exist two sets $I_1, I_2 \in \mathcal{I}$ that are smaller than I such that $I = I_1 \cup I_2$. For any $y \in F^*$, we have $y_{\{I_1,I_2\}} = y_{\{I_1\}} y_{\{I_2\}}$ by definition. By induction,

$$y_{\{I_1\}} = \prod_{a \in I_1} y_{\{\{a\}\}} \quad \text{and} \quad y_{\{I_2\}} = \prod_{a \in I_2} y_{\{\{a\}\}} \,.$$

Finally, the equations in Y enforce $y_{\{I_1,I_2\}} = y_{\{I_1 \cup I_2\}}$. Connecting all this, we derive

$$y_{\{I\}} = y_{\{I_1 \cup I_2\}} = y_{\{I_1,I_2\}} = y_{\{I_1\}} y_{\{I_2\}} = \prod_{a \in I_1} y_{\{\{a\}\}} \prod_{a \in I_2} y_{\{\{a\}\}} = \prod_{a \in I} y_{\{\{a\}\}} \,. \qquad \square$$

If $\mathcal{I}$ is not reducible, then Theorem 3.2 does not hold in general. To see this, consider the unconstrained case, and assume that there is an element $I \in \mathcal{I}$ with $|I| \geq 2$ such that $I \neq I_1 \cup I_2$ for all $I_1, I_2 \in \mathcal{I} \setminus \{I\}$. In other words, $I_1 \cup I_2 = I$ implies $I_1 = I$ or $I_2 = I$. Define $y \in \mathbb{R}^{\mathcal{I}^*}$ by

$$y_{\{I_1,I_2\}} = \begin{cases} 1 & \text{if } I_1 \cup I_2 \subseteq I \text{ but } I_1 \cup I_2 \neq I, \\ 0 & \text{otherwise.} \end{cases}$$

Then $y \in F^* \cap Y$, but the unique preimage $\psi^{-1}(y)$ of y in $\mathbb{R}^{\mathcal{I}}$ is not contained in F, as $\psi^{-1}(y)_I = y_{\{I\}} = 0$, while

$$\prod_{a \in I} \psi^{-1}(y)_{\{a\}} = \prod_{a \in I} y_{\{\{a\}\}} = 1 \,.$$

The latter is true because $|I| \geq 2$ implies that each $\{a\}$ is a proper subset of I.

THEOREM 3.3. *The polytope $P^* \cap Y$ is a face of P^* and thus integer.*

Proof. Consider the linear subspace of $\mathbb{R}^{\mathcal{I}^*}$ given as

$$Y' = \left\{ y \in \mathbb{R}^{\mathcal{I}^*} \mid y_{\{I \cup J\}} = y_{\{I, I \cup J\}} \text{ for all } I, J \in \mathcal{I} \text{ with } I \cup J \in \mathcal{I} \right\}.$$

As the equations defining Y' form a subset of those defining Y, we have $Y \subseteq Y'$. In particular, we have $P^* \cap Y = P^* \cap Y' \cap Y$. We will first intersect P^* with Y' and then with Y and show that in both steps the added equations are induced by valid inequalities for the corresponding polytope, so that we cut out a face in each step.

For the first step, notice that the inequality $y_{\{I \cup J\}} \geq y_{\{I, I \cup J\}}$ is valid for P^* for all $I, J \in \mathcal{I}$ with $I \cup J \in \mathcal{I}$. Hence $P^* \cap Y'$ is a face of P^* and thus integer. Now we claim that for $P^* \cap Y'$ the inequality $y_{\{I \cup J\}} \leq y_{\{I,J\}}$ is valid. As $P^* \cap Y'$ is integer, we have to show this only for integer vectors in this polytope. For these, we have $y_{\{I \cup J\}} = y_{\{I, I \cup J\}} = y_{\{I\}} y_{\{I \cup J\}}$ by definition, so that either $y_{\{I \cup J\}} = 0$ or $y_{\{I\}} = 1$. The same argument for J instead of I yields $y_{\{I \cup J\}} = 0$ or $y_{\{J\}} = 1$. Thus, if $y_{\{I \cup J\}} = 1$, we have $y_{\{I\}} = y_{\{J\}} = 1$, and hence $y_{\{I,J\}} = y_{\{I\}} y_{\{J\}} = 1$. We derive $y_{\{I \cup J\}} \leq y_{\{I,J\}}$. ∎

COROLLARY 3.4. *If $\mathcal{I}$ is reducible, then P is isomorphic to a face of P^* via ψ.*

To conclude this section, we consider the basic SDP relaxation of the quadratic problem (3.1). For any set B, define $\mathcal{P}_k(B) = \{B' \subseteq B \mid |B'| \leq k\}$. For $\mathcal{B} \subseteq 2^B$, let $M_{\mathcal{B}}(y)$ denote the restriction of the moment matrix of y over B to rows and columns indexed over $\mathcal{B}$, and let $*$ denote the shift operator [18, 20]. With these definitions, we have $\mathcal{I}^* \subseteq \mathcal{P}_2(\mathcal{I})$, and the basic SDP relaxation of P^* is

$$\left\{ y \in \mathbb{R}^{\mathcal{P}_2(\mathcal{I})} \mid M_{\mathcal{P}_1(\mathcal{I})}(y) \succeq 0,\ M_{\mathcal{P}_1(\mathcal{I})}(h_i^* * y) \succeq 0 \text{ for } i = 1, \ldots, r,\ y_\emptyset = 1 \right\}. \tag{3.2}$$

Note that $M_{\mathcal{P}_1(\mathcal{I})}(y)_{\{I\},\{J\}} = y_{\{I,J\}}$. Under the linear map ψ, the variable $y_{\{I,J\}}$ corresponds to $z_{I \cup J}$ in (2.2), so that (3.2) is equivalent to

$$\left\{ z \in \mathbb{R}^{\mathcal{I}_2 \cup \{\emptyset\}} \mid M_{\mathcal{I} \cup \{\emptyset\}}(z) \succeq 0,\ M_{\mathcal{I} \cup \{\emptyset\}}(h_i * z) \succeq 0 \text{ for } i = 1, \ldots, r,\ z_\emptyset = 1 \right\}, \tag{3.3}$$

where $\mathcal{I}_2 = \{I \cup J \mid I, J \in \mathcal{I}\}$. Projecting (3.3) to $\mathbb{R}^{\mathcal{I}}$ yields an SDP relaxation of P. Moreover, it is easy to see that the results of this section remain valid if we remove each element $\{I\}$ from $\mathcal{I}^*$ for which I is maximal in $\mathcal{I}$. Then instead of (3.3) we get

$$\left\{ z \in \mathbb{R}^{\mathcal{I}'_2 \cup \{\emptyset\}} \mid M_{\mathcal{I}' \cup \{\emptyset\}}(z) \succeq 0,\ M_{\mathcal{I}' \cup \{\emptyset\}}(h_i * z) \succeq 0 \text{ for } i = 1, \ldots, r,\ z_\emptyset = 1 \right\}, \tag{3.4}$$

where $\mathcal{I}'$ is the set of nonmaximal elements of $\mathcal{I}$ and $\mathcal{I}'_2 = \{I \cup J \mid I, J \in \mathcal{I}'\}$. Now reducibility of $\mathcal{I}$ implies $\mathcal{I} \subseteq \mathcal{I}'_2$, so that (3.4) in this case still induces a relaxation of P, defined by an even smaller number of variables. If the original problem (2.1) is already quadratic, then (3.4) is just the standard SDP relaxation of (2.1).

If $\mathcal{I}$ is reducible, then by Theorem 3.2 the integer points in (3.4) correspond bijectively to the solutions of (2.2). On the other hand, if $\mathcal{I}$ is not reducible, then the example given above shows that in general we can have $M_{\mathcal{I} \cup \{\emptyset\}}(z) \succeq 0$ even if z is integer and is infeasible for (2.2).

Using the framework of Laurent [20], the relaxation (3.4) can be compared to the ones presented by Sherali and Adams [27] and Lasserre [18]: in all three cases, the semidefiniteness of the (full) moment matrices is relaxed by requiring this property only for certain submatrices. In particular, one finds that (3.4) contains the tth Lasserre iterate if all monomials in $\mathcal{I}$ have degree at most $t + 2$ but not earlier in

general. On the other hand, our relaxation is not usually contained in any Lasserre iterate, as we restrict ourselves to very few rows and columns of the moment matrices, corresponding to our aim to keep the number of variables as small as possible.

3.2. The reduction procedure. We showed that the polytope P is isomorphic to a face of a polytope P^* corresponding to a quadratic instance if $\mathcal{I}$ is reducible. In particular, in this case, the separation problem for P reduces to the separation problem for P^*; i.e., the general separation problem reduces to the one for the quadratic case. More precisely, assume that a separation algorithm $\mathcal{A}$ is given for P^*; then we obtain the following induced separation algorithm for P.

Separation of a vector $\overline{z} \in \mathbb{R}^{\mathcal{I}}$ from P.
Compute the vector $\overline{y}^* = \psi(\overline{z}) \in \mathbb{R}^{\mathcal{I}^*}$
Apply the separation algorithm $\mathcal{A}$ to $\overline{y}^*$
if a violated constraint Γ is found
then
  Replace every variable $y_{\{I,J\}}$ in Γ by $z_{I \cup J}$
  Return the resulting constraint
else output "no cutting plane found"

By Lemma 3.1, any constraint returned by this algorithm is valid for P but violated by $\overline{z}$. In particular, the algorithm is correct whenever the answer is positive, i.e., whenever a cutting plane is found. On the other hand, its effectiveness, i.e., the probability of finding a violated cutting plane if one exists, obviously depends on the effectiveness of the underlying separation algorithm $\mathcal{A}$. According to Corollary 3.4, the former algorithm is an exact separation algorithm for P if the same is true for $\mathcal{A}$ and $\mathcal{I}$ is reducible. The reducibility of $\mathcal{I}$ can be obtained artificially by adding new zero-weight sets to $\mathcal{I}$ in a preprocessing step (see section 3.3).

The general idea supporting our approach is thus to solve all LP relaxations in the original variable space (including variables needed for making $\mathcal{I}$ reducible) and to move to the higher dimension of P^* only during separation. The advantage of dealing with P^* instead of P in the separation phase is that the former belongs to the smaller class of polytopes corresponding to quadratic problems, so that there is more hope in finding a good polyhedral description of P^*. For instance, in the unconstrained case, P^* is isomorphic to a cut polytope, while in the general case, the polytope P^* corresponds to a max-cut problem with additional constraints. For cut polytopes, several classes of cutting planes and sophisticated separation techniques are known, while nothing similar is at hand for the more general polytope P, even in the unconstrained case. The practical benefit of this approach is underlined by the results of a first experimental evaluation, presented in section 4.3.

3.3. How to make the instance reducible. As pointed out, our separation approach can be expected to have a good performance only if the given set $\mathcal{I}$ is reducible, even if the validity of the induced inequalities does not require reducibility of $\mathcal{I}$. Obviously, every set $\mathcal{I}$ can be made reducible by adding new elements, e.g., by adding all subsets of elements in $\mathcal{I}$. Unfortunately, the added elements have to be represented by additional variables, so that the problem size increases. This gives rise to the question of how many elements we have to add to $\mathcal{I}$ and how we can find a possibly small set of new elements algorithmically.

One can show that determining a smallest set of additional variables sufficient to make $\mathcal{I}$ reducible is an NP-hard problem, which remains true even for degree $d = 3$. This follows from the NP-hardness of finding an optimal replacement strategy

in the direct reduction approach [5] (see section 3.5). Hence we have to resort to heuristic methods. In our implementation, we use the following straightforward greedy approach: first, add all singletons $\{a\}$ to $\mathcal{I}$ if necessary. Then, as long as $\mathcal{I}$ is not reducible, determine two distinct variables $a_1, a_2 \in A$ such that the cardinality of

$$\mathcal{P}(a_1, a_2) = \Big\{ I \in \mathcal{I} \mid a_1, a_2 \in I \text{ and } I \neq I_1 \cup I_2 \text{ for all } I_1, I_2 \in \mathcal{I} \setminus \{I\} \Big\}$$

is maximal. Now add the sets $\{a_1, a_2\}$ and $I \setminus \{a_1, a_2\}$ to $\mathcal{I}$ for all $I \in \mathcal{P}(a_1, a_2)$. In the worst case, i.e., if all $I \in \mathcal{I}$ are pairwise disjoint and no singletons exist in $\mathcal{I}$, the number of new terms produced by this algorithm is $n + \sum_{I \in \mathcal{I}} (|I| - 2)$ and thus bounded by $n + (d-2)m$. For hard instances, however, the sets $I \in \mathcal{I}$ typically overlap to a significant extent; thus in practice the average number of new variables can be expected to be much smaller than $n + (d-2)m$.

3.4. How to define the separation instance. The set $\mathcal{I}^*$ as defined in section 3.1 has, in general, quadratic size in m. However, assuming that $\mathcal{I}$ is reducible, it is readily checked that all results of section 3.1 remain true if $\mathcal{I}^*$ is constructed in a sparser way as follows: for every $I \in \mathcal{I}$, choose two smaller elements $I_1, I_2 \in \mathcal{I}$ with $I = I_1 \cup I_2$ and add $\{I_1, I_2\}$, $\{I_1, I\}$, and $\{I_2, I\}$ to $\mathcal{I}^*$. In other words, we have to model only a single representation of I as a union of smaller elements in $\mathcal{I}$. Using this construction, the size of $\mathcal{I}^*$ is at most four times the size of $\mathcal{I}$. A further improvement is obtained by omitting any singleton $\{I\}$ for which I is maximal in $\mathcal{I}$.

In summary, these modifications yield a smaller instance $\mathcal{I}^*$ than the one proposed in section 3.1. For example, if the original instance $\mathcal{I}$ was already a quadratic instance, we now have $\mathcal{I} = \mathcal{I}^*$. So it might seem more appealing to use this *sparse* definition of $\mathcal{I}^*$. However, as for the quadratic case, a perfect separation algorithm is not available; this is not always preferable from a computational point of view, because the original *dense* definition leads to tighter LP relaxations in general.

Experimentally, we found that using the dense definition often leads to better results. This, however, might depend on the separation algorithm used for the underlying quadratic problem. It should be worthwhile to search for criteria that allow us to decide whether or not a given pair of monomials should be added to $\mathcal{I}^*$, i.e., whether it is likely that the resulting improvement of the relaxation justifies the new variable. One could even take this decision dynamically, at the beginning of each separation step, based on the current fractional solution $\overline{z}$ to be separated. We have not included this in our implementation yet but plan to examine such strategies in future work.

3.5. Connection to the direct reduction approach. The direct reduction approach, proposed by Rosenberg [26] for the unconstrained case, can be easily extended to arbitrary polynomial zero-one optimization problems as in (2.1). It proceeds as follows: choose two variables x_a and x_b and replace their product $x_a x_b$ by a new variable x_c wherever it appears in a monomial of any of the polynomials g_i, $i = 0, \ldots, r$. Add the quadratic term $-M(x_a x_b - 2x_a x_c - 2x_b x_c + 3x_c)$ with $M \gg 0$ to the objective function in order to enforce $x_c = x_a x_b$ for every optimal solution. Iterate this replacement until all monomials have been reduced to products of at most two variables. In this way, one obtains a quadratic problem instance; denote the corresponding monomial set by $\mathcal{I}'$. In the unconstrained case, the problem is now equivalent to the max-cut problem on some graph G'.

The size of $\mathcal{I}'$ strongly depends on the strategy of choosing the next pair of variables to replace. The connection to our approach is that, whatever strategy is

chosen, there is a corresponding strategy for making $\mathcal{I}$ reducible, such that the graph for which we have to separate from the cut polytope is just G'—if we use the sparse definition of $\mathcal{I}^*$. The opposite is also true: every strategy for making $\mathcal{I}$ reducible induces a replacement strategy such that the resulting graphs agree. Thus solving $\mathcal{I}'$ is equivalent to solving the max-cut problem on exactly the same graph that in our algorithm is used for separation. The main advantage of our approach is that all other parts of the algorithm are carried out on the instance arising from making $\mathcal{I}$ reducible, which is much smaller than $\mathcal{I}^*$. Experimental evidence of this claim is provided in section 4.3. Furthermore, we can also choose the dense definition of $\mathcal{I}^*$, which yields better results in most cases (see section 3.4). Another advantage of our approach is that there is no need for "big M" techniques to reduce the problem, which typically introduces numerical difficulties in the resulting problem.

4. The unconstrained case. In the remainder of this paper, we restrict ourselves to the unconstrained case; i.e., we consider the problem

$$\max\left\{g_0(x) \mid x \in \{0,1\}^A\right\} \tag{4.1}$$

and its standard linearization

$$\begin{array}{lrcl} \max & h_0(z) & & \\ \text{s.t.} & z_I & = & \prod_{a\in I} z_{\{a\}} \quad \text{for all } I \in \mathcal{I}, \\ & z & \in & \{0,1\}^{\mathcal{I}} \end{array} \tag{4.2}$$

with the corresponding polytope P and set of nonconstant monomials $\mathcal{I}$. Observe that P depends only on $\mathcal{I}$ now and that the feasible solutions correspond to all subsets $S \subseteq A$. From now on, the characteristic vector for S in P is denoted by $\overline{\chi}^S$; i.e., we define $\overline{\chi}^S \in \mathbb{R}^{\mathcal{I}}$ by

$$\overline{\chi}^S_I = \begin{cases} 1 & \text{if } I \subseteq S, \\ 0 & \text{otherwise.} \end{cases}$$

Problem (3.1) in section 3.1 now becomes (the linearization of) an unconstrained quadratic zero-one optimization problem. This implies that the corresponding polytope P^* is a *boolean quadric polytope*. These polytopes have been studied extensively by Padberg [23]. Later it was shown by De Simone that every boolean quadric polytope is in fact isomorphic to a cut polytope [8]. In our case, the graph corresponding to this cut polytope is just $(\mathcal{I}, \mathcal{I}^* \setminus \mathcal{I})$ extended by a root node r, i.e., a new node adjacent to all other nodes in $\mathcal{I}$. The transformation of variables between the two formulations yielding the isomorphism is

$$z_J \mapsto \begin{cases} x_{(r,I)} & \text{if } J = \{I\}, \\ \frac{1}{2}\left(x_{(r,I_1)} + x_{(r,I_2)} - x_{(I_1,I_2)}\right) & \text{if } J = \{I_1, I_2\}, \end{cases}$$

$$x_{(I_1,I_2)} \mapsto \begin{cases} z_{\{I_2\}} & \text{if } I_1 = r, \\ z_{\{I_1\}} & \text{if } I_2 = r, \\ z_{\{I_1\}} + z_{\{I_2\}} - 2z_{\{I_1,I_2\}} & \text{otherwise.} \end{cases}$$

Here $x_{(I_1,I_2)}$ denotes the variable in the maximum cut formulation corresponding to an edge (I_1, I_2) in the graph $(\mathcal{I}, \mathcal{I}^* \setminus \mathcal{I})$. From this, we derive the following.

COROLLARY 4.1. *The polytope P is isomorphic to a face of a cut polytope.*

This is a direct consequence of Corollary 3.4. In particular, the separation problem for our original polytope P reduces to a separation problem for a cut polytope in the way explained in section 3.2. As the cut polytope is one of the best-studied objects in polyhedral combinatorics, we can fall back upon a large number of known classes of cutting planes and, in particular, sophisticated separation techniques for the cut polytope in order to solve the unconstrained polynomial problem (4.1) (see, e.g., [21]).

Before reporting results of a straightforward implementation of this approach, we examine P in two special cases with respect to the monomial set $\mathcal{I}$, in which P has some useful properties: the case that $\mathcal{I}$ is closed under taking supersets, and the opposite case that $\mathcal{I}$ is closed under taking subsets.

4.1. Upward completeness. Throughout this section, we assume that $\mathcal{I}$ is closed under taking supersets, i.e., that

$$I \in \mathcal{I},\ J \in 2^A,\ I \subseteq J \quad \Longrightarrow \quad J \in \mathcal{I}\ .$$

In other words, the product of every monomial $I \in \mathcal{I}$ with every variable x_a, $a \in A$, is again a monomial in $\mathcal{I}$. In this case, we have the following.

THEOREM 4.2. *The polytope P is isomorphic to the simplex of dimension m.*

Proof. We first claim that P is given by the linear inequalities

$$(-1)^{|I|} \sum_{J \supseteq I} (-1)^{|J|} z_J \geq 0 \tag{4.3}$$

for every $I \in \mathcal{I}$ and

$$\sum_{I \in \mathcal{I}} (-1)^{|I|} \sum_{J \supseteq I} (-1)^{|J|} z_J \leq 1\ . \tag{4.4}$$

These constraints are valid, as for integer vectors in P we have

$$(-1)^{|I|} \sum_{J \supseteq I} (-1)^{|J|} z_J = \prod_{a \in I} x_a \prod_{a \in A \setminus I} (1 - x_a)\ .$$

Now an isomorphism between the simplex of dimension m and the polytope P' given by (4.3) and (4.4) is induced by the variable transformations

$$z_I \mapsto \sum_{J \supseteq I} z'_J \quad \text{and} \quad z'_I \mapsto (-1)^{|I|} \sum_{J \supseteq I} (-1)^{|J|} z_J \quad \text{for all } I \in \mathcal{I}\ .$$

These transformations are well defined because $\mathcal{I}$ is closed under taking supersets. Moreover, they are inverse to each other, as one can check with some patience.

Under the given transformation, the inequality (4.3) becomes $z'_I \geq 0$, while (4.4) is transformed into $\sum_{I \in \mathcal{I}} z'_I \leq 1$. Hence P' is isomorphic to the simplex of dimension m. Furthermore, it is readily checked that under this transformation all vertices of the simplex correspond to characteristic vectors $\overline{\chi}^S$ for suitable $S \subseteq A$. In particular, we derive $P' \subseteq P$. The validity of (4.3) and (4.4) implies $P = P'$. □

Note that the transformation used in the proof of Theorem 4.2 is a multiplication with the so-called *zeta matrix*; see, e.g., section 3.1 in [20]. The constraints (4.3) and (4.4) are sometimes called *bound-factor product constraints*.

The condition that $\mathcal{I}$ be closed under taking supersets is of purely theoretical interest, as usually it will not hold for practical instances. Moreover, the number

of elements that would have to be added to $\mathcal{I}$ in order to achieve this property is exponential in general. However, the complete variable set $2^A \setminus \{\emptyset\}$ meets this condition, and it is easy to see that every polytope P is a projection of the polytope $\overline{P}$ corresponding to $2^A \setminus \{\emptyset\}$, which is nothing but the n-th Sherali-Adams iterate [27] of the cube $[0,1]^A$. Theorem 4.2 states that $\overline{P}$ is a simplex of dimension $2^n - 1$, implying that P is full-dimensional for every $\mathcal{I}$. However, in general, more interesting properties of $\overline{P}$ do not carry over via this projection.

4.2. Downward completeness. In the previous section, we examined the special case where $\mathcal{I}$ is closed under taking supersets. In this case, the polytope P turned out to have a very simple structure. In this section, we assume, on the contrary, that $\mathcal{I}$ is closed under taking nonempty subsets; i.e., we suppose

$$I \in \mathcal{I},\ J \in 2^A \setminus \{\emptyset\},\ I \supseteq J \quad \Longrightarrow \quad J \in \mathcal{I}.$$

This case is practically much more relevant than the one considered previously. For instance, it contains all quadratic instances and hence the max-cut problem as a special case. However, in this case the polytope P shares some interesting properties with the cut polytope of which it can consequently be considered as the natural generalization.

A well-known property of the cut polytopes is their symmetry: for each two vertices, there is an automorphism of the polytope, called *switching*, that maps the first vertex to the second (see, e.g., [9]). We claim that this is true in general if $\mathcal{I}$ is closed under taking nonempty subsets. Moreover, switching has a most natural interpretation: switching with respect to $S \subseteq A$ amounts to switching the value of each variable x_a with $a \in S$.

THEOREM 4.3. *Let P be the convex hull of feasible solutions of* (4.2)*, and let $\{a\} \in \mathcal{I}$ for all $a \in A$. Then for each $S \subseteq A$, a switching π_S of the vertices of P is given by $\pi_S(\overline{\chi}^T) = \overline{\chi}^{T \triangle S}$. Furthermore, the following statements are equivalent:*

(a) *The set $\mathcal{I}$ is closed under taking nonempty subsets.*

(b) *Each switching π_S corresponds to an automorphism of P.*

Proof. Since $\{a\} \in \mathcal{I}$ for all $a \in A$, the characteristic vectors $\overline{\chi}^T$ for $T \subseteq A$ are pairwise distinct, so that the permutation π_S is well defined. To show that (a) implies (b), consider the affine map $\overline{\pi}_S \colon \mathbb{R}^{\mathcal{I}} \to \mathbb{R}^{\mathcal{I}}$, defined componentwise by

$$\overline{\pi}_S(z)_I = (-1)^{|I \setminus S|} \sum_{I \setminus S \subseteq J \subseteq I} (-1)^{|J|} z_J$$

for $z \in \mathbb{R}^{\mathcal{I}}$ and $I \in \mathcal{I}$. Here we set $z_\emptyset = 1$ for all $z \in \mathbb{R}^{\mathcal{I}}$. This map is well defined, as $\mathcal{I}$ is closed under taking subsets. Moreover, as observed in section 4.1, it brings $0-1$ points to $0-1$ points; thus it is an automorphism of P. It remains to show that $\overline{\pi}_S(\overline{\chi}^T) = \overline{\chi}^{T \triangle S}$ for all $T \subseteq A$, since this implies that $\overline{\pi}_S$ induces a switching of the vertices of P and hence an automorphism π_S of P. Indeed, arguing componentwise, we have

$$\begin{aligned}
\overline{\chi}_I^{T \triangle S} &= \prod_{a \in I} \overline{\chi}^{\{a\}} \cdot \overline{\chi}^{T \triangle S} = \prod_{a \in I \setminus S} \overline{\chi}^{\{a\}} \cdot \overline{\chi}^T \prod_{a \in I \cap S} \left(1 - \overline{\chi}^{\{a\}} \cdot \overline{\chi}^T\right) \\
&= (-1)^{|I \setminus S|} \sum_{I \setminus S \subseteq J \subseteq I} (-1)^{|J|} \prod_{a \in J} \overline{\chi}^{\{a\}} \cdot \overline{\chi}^T = (-1)^{|I \setminus S|} \sum_{I \setminus S \subseteq J \subseteq I} (-1)^{|J|} \overline{\chi}_J^T \\
&= \overline{\pi}_S(\overline{\chi}^T)_I
\end{aligned}$$

for all $I \in \mathcal{I}$. This completes the first part of the proof.

Now let (a) be violated. Then there are sets $S \subset T \subseteq A$ with $T \in \mathcal{I}$ but $T \setminus S \notin \mathcal{I}$. One may assume that S is maximal, so that all nonempty proper subsets of $T \setminus S$ belong to $\mathcal{I}$. We claim that the $2^{|T \setminus S|}$ vectors $\{\overline{\chi}^R \mid R \subseteq T \setminus S\}$ are affinely dependent in $\mathbb{R}^{\mathcal{I}}$, while their images under π_S are not. This implies that π_S is not induced by any automorphism of P.

Indeed, these vectors are 0 in all dimensions I with $I \not\subseteq T \setminus S$. As $\emptyset$ and $T \setminus S$ do not belong to $\mathcal{I}$, the corresponding dimensions do not exist in $\mathbb{R}^{\mathcal{I}}$. Therefore, the $2^{|T \setminus S|}$ vectors can differ only in the remaining $2^{|T \setminus S|} - 2$ dimensions and are thus affinely dependent. On the other hand, their images $\{\overline{\chi}^{R \cup S} \mid R \subseteq T \setminus S\}$ under π_S are affinely independent, as can be checked easily by inspecting the dimensions I with $\emptyset \neq I \subset T \setminus S$ and $I = T$, which by construction all exist in $\mathbb{R}^{\mathcal{I}}$. □

COROLLARY 4.4. *Any vertex of the polytope P can be mapped to any other vertex by a single automorphism π_S. In particular, all affine cones pointed at the vertices of P are isomorphic.*

COROLLARY 4.5. *Let $a \in \mathbb{R}^{\mathcal{I}}$ and $\alpha \in \mathbb{R}$. Then the inequality $a \cdot y \leq \alpha$ is valid for P if and only if the inequality $a \cdot \pi_S(y) \leq \alpha$ is valid for P. The former is facet-inducing for P if and only if the latter is.*

4.3. Experimental evaluation. In this section, we report preliminary experimental results obtained with a straightforward implementation of the ideas presented in section 3, applied to the unconstrained case. They are meant only to give a first indication of the practical performance and usability to be expected from the presented approach. Note that the entire implementation is a simple task if a separation procedure for the cut polytope is ready to hand; one has to only code the trivial transformations explained in section 3.2 and embed everything into a branch-and-cut framework. For this, we used ABACUS [17] in combination with CPLEX [6]. For the max-cut separation, we used only *cycle inequalities* here, with the well-known exact separation algorithm [4]. We did not apply any tailing off strategy; for enumeration we used the depth first approach.

We experimented with randomly generated instances of a given degree and density. In the following, we focus on instances with a small degree. For given degree d, number of variables n, and number of monomials m, the instances were obtained by randomly choosing m subsets of $\{1, \ldots, n\}$ with d elements each. Objective function coefficients were chosen randomly from the reals in the interval $[-1, 1]$.

For each selected set of parameters d, n, m, we solved 10 random instances on a Pentium 4 processor with 2.8 GHz and 1 GB of main memory. The results are displayed in Table 4.1. We report average figures for the total runtime in cpu-seconds needed to solve the instance to optimality (Time) and for the number of nodes in the enumeration tree (Subs). For comparison, we state these data for the CPLEX MIP-solver [6] applied to the basic relaxation given by (2.3) and (2.4), for the direct reduction approach discussed in section 3.5, and for the algorithm presented in this paper. In some cases, CPLEX MIP or direct reduction could not solve all instances within 24 cpu-hours. For these cases, we state lower bounds on both the runtime and the number of subproblems.

In order to obtain comparable figures, we used the same implementation framework for both the direct reduction and our algorithm; i.e., after reducing the objective function we applied our own implementation to the resulting quadratic instance. In Table 4.2, we show the average node degree of the corresponding separation graphs and the quality of the upper bound at the root node.

TABLE 4.1
Results for small-degree instances.

Instances			CPLEX MIP		Direct reduction		Our algorithm	
d	n	m	Time	Subs	Time	Subs	Time	Subs
3	200	400	8.20	374.3	3.47	1.2	2.07	1.2
3	200	500	1088.08	61588.4	156.27	24.4	69.11	16.8
3	200	600	>49859.67	>2104977.5	12935.19	1259.4	3107.30	549.8
3	400	700	15.06	328.5	19.59	2.0	16.81	2.6
3	400	800	2080.75	40455.9	505.06	35.6	147.32	14.8
3	400	900	>51341.72	>706246.6	>11597.68	>488.4	6517.69	416.6
3	600	1000	75.84	1010.0	31.62	1.2	21.11	1.2
3	600	1100	712.00	5695.1	443.83	13.8	248.53	10.4
3	600	1200	>43309.47	>369578.1	>18989.30	>425.0	11202.83	369.2
4	200	250	2.58	115.3	10.14	3.0	8.25	3.6
4	200	300	63.77	3550.0	252.86	43.0	105.80	27.8
4	200	350	1483.77	77808.8	2250.22	227.0	403.94	72.0
4	400	450	5.39	89.1	40.76	3.6	23.03	3.0
4	400	500	17.11	372.9	486.20	36.4	84.31	7.6
4	400	550	1416.09	26638.5	11217.64	443.8	737.08	43.2
4	600	650	6.41	45.9	31.38	1.4	21.60	1.6
4	600	700	43.66	484.1	433.53	12.8	164.25	6.2
4	600	750	489.12	4675.4	>9670.97	>211.8	2194.87	65.6

As we used only cycle inequalities for separation, the given bounds at the root node correspond to the relaxation of P induced by the cycle relaxation of P^*. The results in Table 4.1 show that these bounds are rather tight for the examined instances: few LPs and subproblems have to be solved in general. On the other hand, a single separation phase takes a lot of time—for some instances, separation consumed more than 50 % of the total runtime. To a certain extent, this is due to the nature of our approach, as the separation instance is considerably larger than the LPs to be solved. Nevertheless, there is a lot of room for improvement at this point, e.g., by using faster separation algorithms such as the spanning-tree heuristic for separation of cycle inequalities; see, for example, the recent survey [21].

For increasing degree d, the relaxation for P induced by the cycle relaxation of P^* will become weaker in general. Furthermore, the number of additional variables for making $\mathcal{I}$ reducible will increase. We thus expect runtimes to increase with d. On the other hand, higher degrees mean a stronger interrelation between different monomials, making branching more efficient in general, as setting a monomial to 0 or 1 has a greater effect in this case. Similar reasoning applies to the density m/n: for sparse instances, the LP bounds are much better than for dense ones, while branching has a greater effect for the latter instances. Therefore, our approach outperforms CPLEX MIP much more clearly for instances of small degree and density in general.

All instances considered so far have a small degree. To examine the effect of higher-degree monomials, we created another test set containing higher-degree instances; to keep density low at the same time, we used an exponential distribution for this. More precisely, each of the m monomials was chosen as follows: first, its degree $d \in \{2, \dots, n\}$ was determined randomly, where the probability of choosing degree $d < n$ was 2^{1-d}. Then the monomial was picked randomly from the set of all possible monomials of the chosen degree. Again, we created 10 instances for each combination of m and n. This generation method yields instances of high degree but

TABLE 4.2
Root bounds for small-degree instances.

Instances			Dense	
d	n	m	Deg	Bnd
3	200	400	5.3	0.0 %
3	200	500	5.7	0.8 %
3	200	600	6.0	3.6 %
3	400	700	5.1	0.0 %
3	400	800	5.3	0.3 %
3	400	900	5.5	1.0 %
3	600	1000	5.0	0.0 %
3	600	1100	5.2	0.1 %
3	600	1200	5.3	0.6 %
4	200	250	4.9	0.1 %
4	200	300	5.2	1.1 %
4	200	350	5.4	2.6 %
4	400	450	4.8	0.1 %
4	400	500	5.0	0.2 %
4	400	550	5.1	0.6 %
4	600	650	4.8	0.0 %
4	600	700	4.9	0.1 %
4	600	750	5.0	0.4 %

with many lower-degree monomials, which is a situation likely to arise in practical applications. Note that the expected degree of a monomial in these instances is almost three. Nevertheless, results were much better than for instances with all monomials of degree three (see Tables 4.3 and 4.4).

TABLE 4.3
Results for sparse higher-degree instances.

Instances			CPLEX MIP		Direct reduction		Our algorithm	
d	n	m	Time	Subs	Time	Subs	Time	Subs
11.5	200	500	5.16	297.6	4.03	1.0	2.80	1.0
12.4	200	600	408.65	33585.8	859.44	121.0	139.53	24.0
12.2	200	700	2676.27	172987.4	384.23	35.0	342.52	43.8
12.4	200	800	>42411.65	>2004332.2	9740.83	566.0	2359.17	207.2
12.1	400	800	5.92	42.8	9.67	1.4	5.55	1.4
12.1	400	900	32.11	882.5	39.27	2.4	28.28	2.4
12.1	400	1000	432.43	9431.9	293.76	11.8	95.35	4.4
12.3	400	1100	>37274.49	>891109.4	7161.89	239.8	3029.73	141.4
13.3	600	1100	35.58	946.5	33.58	1.2	25.40	1.2
13.9	600	1200	130.92	2769.4	311.97	9.4	171.38	6.6
13.7	600	1300	540.41	9325.6	336.80	7.4	194.50	5.2
14.1	600	1400	>12080.52	>185146.6	3186.61	57.6	1751.48	42.2

As pointed out, the results of Tables 4.1 and 4.3 were obtained by a vanilla implementation. Runtimes should decrease considerably by using faster separation routines, further classes of cutting planes, an adequate branching strategy, and so on. Implementing separation routines designed for denser graphs should improve runtimes considerably. As obvious from Tables 4.1 and 4.3, denser separation graphs are caused

TABLE 4.4
Root bounds for sparse higher-degree instances.

Instances			Dense	
d	n	m	Deg	Bnd
11.5	200	500	6.8	0.0 %
12.4	200	600	7.1	0.9 %
12.2	200	700	7.4	0.9 %
12.4	200	800	7.7	2.3 %
12.1	400	800	6.2	0.0 %
12.1	400	900	6.5	0.0 %
12.1	400	1000	6.7	0.1 %
12.3	400	1100	6.8	0.6 %
13.3	600	1100	6.1	0.0 %
13.9	600	1200	6.2	0.0 %
13.7	600	1300	6.4	0.1 %
14.1	600	1400	6.6	0.2 %

by larger ratios m/n and by higher degrees d.

5. Conclusion. We presented a new approach for reducing polynomial zero-one optimization problems to the quadratic case. Unlike other approaches, we do not produce an equivalent quadratic instance directly but use a reduction of the general separation problem to the one for quadratic instances and thus to the separation problem for the cut polytope with additional linear constraints. All other components of a branch-and-cut algorithm such as the solution of the LP relaxations, primal heuristics, or branching can be performed on a much smaller instance that arises from making the original instance reducible.

Experimental results show that our reduction method is a promising approach for solving large unconstrained polynomial optimization problems. Here we make use of the fact that the separation problem for the quadratic case is well studied. Unfortunately, the situation is less comfortable in the presence of constraints. Certain constrained quadratic problems, such as the quadratic assignment problem or the quadratic knapsack problem, have been investigated from a polyhedral point of view; the corresponding separation algorithms can be applied in our approach in order to address the corresponding polynomial problems. However, general constrained polynomial problems are reduced to general constrained quadratic problems, which themselves are hard enough to solve. Our aim was to show that moving from quadratic to polynomial functions does not add much to the hardness of the problem.

Acknowledgment. We would like to thank the two anonymous referees for their helpful comments and suggestions that improved this paper significantly.

REFERENCES

[1] E. BALAS, S. CERIA, AND G. CORNUÉJOLS, *A lift-and-project cutting plane algorithm for mixed* 0–1 *programs*, Math. Programming, 58 (1993), pp. 295–324.
[2] E. BALAS AND J. B. MAZZOLA, *Nonlinear* 0–1 *programming:* I. *Linearization techniques*, Math. Programming, 30 (1984), pp. 1–21.
[3] E. BALAS AND M. PERREGAARD, *Lift-and-project for mixed* 0–1 *programming: Recent progress*, Discrete Appl. Math., 123 (2002), pp. 129–154.
[4] F. BARAHONA AND A. R. MAHJOUB, *On the cut polytope*, Math. Programming, 36 (1986), pp. 157–173.

[5] E. Boros and P. L. Hammer, *Pseudo-boolean optimization*, Discrete Appl. Math., 123 (2002), pp. 155–225.
[6] Cplex 8.1, http://www.ilog.com/products/cplex.
[7] Y. Crama, P. Hansen, and B. Jaumard, *The basic algorithm for pseudo-boolean programming revisited*, Discrete Appl. Math., 29 (1990), pp. 171–185.
[8] C. De Simone, *The cut polytope and the Boolean quadric polytope*, Discrete Math., 79 (1990), pp. 71–75.
[9] M. Deza and M. Laurent, *Geometry of Cuts and Metrics*, Algorithms Combin. 15, Springer-Verlag, Berlin, 1997.
[10] R. Fortet, *Applications de l'algèbre de Boole en recherche opérationelle*, Rev. Française Rech. Opér., 4 (1960), pp. 17–26.
[11] F. Glover and E. Woolsey, *Further reduction of zero-one polynomial programs to zero-one linear programming problems*, Oper. Res., 21 (1973), pp. 156–161.
[12] F. Glover and E. Woolsey, *Note on converting the* 0–1 *polynomial programming problems to zero-one linear programming problems*, Oper. Res., 22 (1974), pp. 180–181.
[13] F. Granot and P. L. Hammer, *On the use of Boolean functions in* 0–1 *programming*, Methods Oper. Res., 12 (1971), pp. 154–184.
[14] P. L. Hammer, ed., *Special issue: Boolean and pseudo-Boolean optimization*, Discrete Appl. Math., 149 (2005), pp. 1–207.
[15] P. L. Hammer, P. Hansen, and B. Simeone, *Roof duality, complementation and persistency in quadratic* 0–1 *optimization*, Math. Programming, 28 (1984), pp. 121–155.
[16] P. L. Hammer, I. Rosenberg, and S. Rudeanu, *Application of discrete linear programming to the minimization of Boolean functions*, Rev. Roumaine Math. Pures Appl., 8 (1963), pp. 459–475.
[17] M. Jünger and S. Thienel, *The ABACUS system for branch-and-cut-and-price-algorithms in integer programming and combinatorial optimization*, Software: Practice & Experience, 30 (2000), pp. 1325–1352.
[18] J. B. Lasserre, *Semidefinite programming vs. LP relaxations for polynomial programming*, Math. Oper. Res., 27 (2002), pp. 347–360.
[19] M. Laurent, *Max-cut problem*, in Annotated Bibliography in Combinatorial Optimization, M. Dell'Amico, F. Maffioli, and S. Martello, eds., Wiley, Chichester, 1997, pp. 241–259.
[20] M. Laurent, *A comparison of the Sherali-Adams, Lovász-Schrijver, and Lasserre-relaxations for* 0–1 *programming*, Math. Oper. Res., 28 (2003), pp. 470–496.
[21] F. Liers, M. Jünger, G. Reinelt, and G. Rinaldi, *Computing exact ground states of hard Ising spin glass problems by branch-and-cut*, in New Optimization Algorithms in Physics, A. K. Hartmann and H. Rieger, eds., Wiley-VCH, Weinheim, Germany, 2004, pp. 47–69.
[22] L. Lovász and A. Schrijver, *Cones of matrices and set-functions and* 0–1 *optimization*, SIAM J. Optim., 1 (1991), pp. 166–190.
[23] M. Padberg, *The boolean quadric polytope: Some characteristics, facets and relatives*, Math. Programming, 45 (1989), pp. 139–172.
[24] P. A. Parillo, *Semidefinite programming relaxations for semialgebraic problems*, Math. Program., 96 (2003), pp. 293–320.
[25] I. G. Rosenberg, 0–1 *optimization and non-linear programming*, Rev. Française Automat. Informat. Recherche Opérationelle, 6 (1972), pp. 95–97.
[26] I. G. Rosenberg, *Reduction of bivalent maximization to the quadratic case*, Cahiers Centre Études Recherche Opér., 17 (1975), pp. 71–74.
[27] H. D. Sherali and W. P. Adams, *A hierarchy of relaxations between the continuous and convex hull representations for zero-one programming problems*, SIAM J. Discrete Math., 3 (1990), pp. 411–430.
[28] L. G. Watters, *Reduction of integer polynomial problems to zero-one linear programming problems*, Oper. Res., 15 (1967), pp. 1171–1174.

SIAM J. OPTIM.
Vol. 18, No. 4, pp. 1414–1435

A HIGH-ORDER PATH-FOLLOWING METHOD FOR LOCATING THE LEAST 2-NORM SOLUTION OF MONOTONE LCPs*

ANHUA LIN†

Abstract. A high-order path-following method is proposed for finding the least 2-norm solution of a monotone LCP. This method follows the regularized central paths introduced in [Y. B. Zhao and D. Li, *SIAM J. Control Optim.*, 40 (2001), pp. 898–924]. By using the analyticity of these paths, we showed the global convergence under the assumption that the LCP has at least one solution and the superlinear rate of local convergence under the assumption that the least 2-norm solution is maximally complementary.

Key words. least 2-norm, monotone LCP, high-order method

AMS subject classifications. 90C33, 90C51, 49M29

DOI. 10.1137/060659752

1. Introduction.

In this paper we are concerned with finding the least 2-norm solution of the following monotone LCP:

$$0 \le x \perp Mx + q \ge 0, \tag{1}$$

where $q \in R^n$ and $M \in R^{n\times n}$ is positive semidefinite. A monotone LCP may have many solutions, but the least 2-norm solution is unique.

The problem of finding the least 2-norm solution to some optimization and complementarity problems has been extensively studied. See [23, 5, 9] for some recent results on linear programs and [18, 7, 21, 22, 24] on complementarity problems. Also, it is well known that the Tikhonov regularization trajectory of a monotone complementarity problem converges to the least 2-norm solution [1].

On the other hand, there is a vast literature on finding a general solution of LCPs. In particular, the local convergence analysis of interior-point-like methods for monotone LCPs has been studied, to name a few, by Wright and Zhang [19], Ye and Anstreicher [20], McShane [10], and Huang, Qi, and Sun [4], under the assumption that there exists a strictly complementary solution, and by Mizuno [11], Potra and Sheng [12], Sturm [17], Stoer, Wechs, and Mizuno [16], Stoer [14], and Zhao and Sun [25] without such an assumption. The results in the last few papers, namely, the second-order predictor-corrector method based on strictly feasible central paths proposed in [17] and the local analysis of high-order methods based on infeasible central paths presented in [16, 14, 25], are the direct motivation for this paper.

In [21, 22] Zhao and Li introduced a class of new paths for general complementarity problems. This class of paths has many nice properties. For example, it needs weaker conditions to be existent and bounded than the normal central path does. For monotone problems, these paths always converge to the least 2-norm solution as long as the problem has at least one solution. However, along with the good properties come some difficulties for algorithm design. The study of this class of paths

*Received by the editors May 12, 2006; accepted for publication (in revised form) May 6, 2007; published electronically January 16, 2008. This work was partially supported by a 2006–2007 Middle Tennessee State University summer and academic year research grant.

http://www.siam.org/journals/siopt/18-4/65975.html

†Department of Mathematical Sciences, Middle Tennessee State University, Murfreesboro, TN 37132 (alin@mtsu.edu).

is still far less mature than that for the usual central path. In [23] a path-following method based on this class of paths was proposed to find the least 2-norm solution of a linear program. But no local convergence analysis was provided for that method. More recently, Zhao and Li proposed a globally and locally superlinearly convergent method based on these paths for P_0 LCP [24]. However, the superlinear convergence was proved under a rather strong assumption that the solution to the LCP is not only unique but also strictly complementary.

The goal of this paper is to use some variant of this class of paths to design a high-order path-following method to locate the least 2-norm solution of a monotone LCP. We will study the analytical property of these paths and use the result to obtain a much better local convergence under a weaker condition than [24].

The paper is organized as follows. In section 2 we introduce the paths that will be used. Then the algorithm will be presented in section 3. Sections 4 and 5 will be devoted to global and local convergence analysis, respectively. Then in section 6 we summarize and point out some future research directions.

We mention the following notation:

(i) For any matrix A, we use $A \succeq 0$ or $0 \preceq A$ to denote that A is square positive semidefinite, and $A \succ 0$ or $0 \prec A$ to denote that A is square positive definite. But A needs not to be symmetric.

(ii) For any vector x, we use $\|x\|$ and $\|x\|_\infty$ to denote the 2-norm and ∞-norm of x, respectively.

(iii) Given a vector x or vector function $f(s)$, we use the corresponding uppercase symbol X or $F(s)$ to denote the natural diagonal matrix generated by x or $f(s)$.

(iv) We use e to denote the all-1 vector with appropriate dimension.

(v) We use R^n_+ (respectively, R^n_{++}) to denote the nonnegative (respectively, positive) orthant in R^n.

2. A class of regularized central paths. In this paper, we consider the following system:

$$\begin{cases} Xy = t^2 w, \\ y = Mx + tx + q, \\ x, y \in R^n_{++}, \end{cases} \tag{2}$$

where $0 < t \in R$ and $0 < w \in R^n$ are parameters. For each fixed $w > 0$, the solutions form a path as t varies in R_{++}.

This class of paths is actually a subclass of the regularized central path proposed in [21] with different parametrization (which is very important for our purpose).

The following theorem about the basic properties of this class of paths can be easily obtained by following the proof of Theorem 2.1 in [23].

THEOREM 2.1.

(i) *For each $(t, w) > 0$, system (2) has a unique solution $(x(t, w), y(t, w)) > 0$.*

(ii) *For any fixed $w > 0$ and any finite number $0 < t_0 < \infty$, the set $\{x(t, w) | 0 < t \le t_0\}$ is bounded if and only if the monotone LCP (1) has at least one solution.*

(iii) *LCP (1) has at least one solution if and only if for any fixed $w > 0$, $x(t, w)$ converges as $t \longrightarrow 0+$; when this is the case, the limit is the unique least 2-norm solution, denoted by x^*.*

Proof. Let $a = w$, $b = 0$, $\theta = \frac{t^2}{1+t^2} \in (0, 1)$, $\phi(\theta) = \sqrt{\frac{\theta}{1-\theta}}$, and $f(x) = Mx + q$.

Then according to Theorem 4.2(a) of [21], system

$$\begin{cases} Xv = \theta a, \\ v = (1-\theta)(Mx + q + \phi(\theta)x), \\ x, v \in R^n_{++} \end{cases} \tag{3}$$

has a unique solution. While by letting $y = \frac{v}{1-\theta}$, it is easy to see that system (3) is equivalent to system (2). So (i) is proved.

For (ii), if $x(t,w)$ is bounded on $0 < t \le t_0$, then any limit point of $x(t,w)$ as $t \longrightarrow 0+$ solves the LCP (1). Conversely, if the LCP (1) has at least one solution, then noticing

$$0 < t \le t_0 \Longleftrightarrow 0 < \theta \le \frac{t_0^2}{1+t_0^2},$$

we have the boundedness of $x(\theta, w)$ and $x(t,w)$ by Theorem 5.1(b) of [21].

Result (iii) is an immediate consequence of Theorem 5.2 of [21]. □

It is easy to show that $(x(t,w), y(t,w))$ is an analytic vector function for $(t,w) > 0$. First we use the fact that for any $(x,y) > 0$ and a P_0 matrix P, $\left[\begin{smallmatrix} Y & X \\ P & -I \end{smallmatrix}\right]$ is nonsingular [6], and noticing that any monotone matrix is a P_0 matrix, we have the following lemma.

LEMMA 2.2. *For positive diagonal matrices $X, Y \in R^{n\times n}$ and any $t \ge 0$, the matrix*

$$J = \begin{bmatrix} Y & X \\ -(M+tI) & I \end{bmatrix}$$

is nonsingular.

THEOREM 2.3. *$(x(t,w), y(t,w))$ is an analytic vector function for $(t,w) > 0$.*

Proof. Let $(\bar{t}, \bar{w})$ be any vector in R^{n+1}_{++}, $(\bar{x}, \bar{y}) = (x(\bar{t}, \bar{w}), y(\bar{t}, \bar{w})) > 0$.

Let $\Phi(x,y,t,w) = \left(\begin{smallmatrix} Xy - t^2 w \\ y - (M+tI)x - q \end{smallmatrix}\right)$. Then $(\bar{x}, \bar{y}, \bar{t}, \bar{w})$ is a solution of $\Phi(x,y,t,w) = 0$.

The Jacobian of $\Phi(x,y,t,w)$ with respect to (x,y) at $(\bar{x}, \bar{y}, \bar{t}, \bar{w})$ is

$$\begin{bmatrix} \bar{Y} & \bar{X} \\ -(M+\bar{t}I) & I \end{bmatrix},$$

which is nonsingular according to Lemma 2.2.

Then from the implicit function theorem and the uniqueness of $(x(t,w), y(t,w))$ for any fixed $(t,w) > 0$, and noticing that $\Phi(x,y,t,w)$ is an analytic vector function of (x,y,t,w), we know that $(x(t,w), y(t,w))$ is an analytic vector function in a neighborhood of $(\bar{t}, \bar{w})$.

Because this is true for all $(\bar{t}, \bar{w}) > 0$, the theorem is then proved. □

Let

$$u(\gamma, t, w) = w + \frac{(\gamma - t)}{1+t}(w - e) = \frac{1+\gamma}{1+t}w + \frac{t-\gamma}{1+t}e. \tag{4}$$

LEMMA 2.4. *If $w > 0$, $t > -1$, and $-1 < \gamma < t + \frac{(1+t)\min\{w_i\}}{1+\min\{w_i\}}$, then $u(\gamma, t, w) > 0$.*

Proof. We consider two cases:

(i) $w_i \geq 1$:

$$\begin{aligned}(1+\gamma)w_i + (t-\gamma) &= w_i + t + \gamma(w_i - 1)\\ &\geq w_i + t - (w_i - 1)\\ &= t + 1\\ &> 0;\end{aligned}$$

(ii) $w_i < 1$:

$$\begin{aligned}(1+\gamma)w_i + (t-\gamma) &= w_i + t + \gamma(w_i - 1)\\ &\geq w_i + t + \left(t + \frac{(1+t)\min\{w_j\}}{1+\min\{w_j\}}\right)(w_i - 1)\\ &= w_i + t + tw_i - t + \frac{(1+t)\min\{w_j\}}{1+\min\{w_j\}}(w_i - 1)\\ &= (1+t)w_i + (1+t)\frac{(w_i - 1)\min\{w_j\}}{1+\min\{w_j\}}\\ &= (1+t)\frac{2w_i\min\{w_j\} + w_i - \min\{w_j\}}{1+\min\{w_j\}}\\ &\geq (1+t)\frac{2w_i\min\{w_j\}}{1+\min\{w_j\}}\\ &> 0.\end{aligned}$$

Therefore $u_i(\gamma,t,w) = \frac{(1+\gamma)w_i+(t-\gamma)}{1+t} > 0$ for all i, and hence $u(\gamma,t,w) > 0$. □

Consider the vector function

$$(\hat{x}(\gamma,t,w), \hat{y}(\gamma,t,w)) = (x(\gamma,u(\gamma,t,w)), y(\gamma,u(\gamma,t,w)))\,.$$

From Theorem 2.3, the definition of $u(\gamma,t,w)$ (4), and Lemma 2.4, it is clear that $(\hat{x}(\gamma,t,w), \hat{y}(\gamma,t,w))$ is a well-defined analytic vector function on the open set

$$\left\{(\gamma,t,w)\,\middle|\, w > 0, t > 0, 0 < \gamma < t + \frac{(1+t)\min\{w_i\}}{1+\min\{w_i\}}\right\}.$$

We will use the partial derivatives $\frac{\partial^l(\hat{x},\hat{y})}{\partial\gamma^l}(t,t,w)$ extensively in our algorithm. Fortunately, they can be easily calculated.

First we introduce the following notation:

(i) $\hat{x}^{(l)}(\gamma,t,w) := \frac{\partial^l\hat{x}}{\partial\gamma^l}(\gamma,t,w),\ \hat{y}^{(l)}(\gamma,t,w) := \frac{\partial^l\hat{y}}{\partial\gamma^l}(\gamma,t,w),\ l = 1,2,\ldots.$

(ii) $\hat{x}^{(0)}(\gamma,t,w) := \hat{x}(\gamma,t,w),\ \hat{y}^{(0)}(\gamma,t,w) := \hat{y}(\gamma,t,w).$

(iii) $(\hat{x}+\hat{y})^{(l)}(\gamma,t,w) := \sum_{i=0}^{l}\binom{l}{i}\hat{X}^{(i)}(\gamma,t,w)\hat{y}^{(l-i)}(\gamma,t,w),\ l = 1,2,\ldots.$

(iv) $(\hat{x}+\hat{y})^{(0)}(\gamma,t,w) := \hat{X}(\gamma,t,w)\hat{y}(\gamma,t,w).$

Since $(\hat{x}(\gamma,t,w), \hat{y}(\gamma,t,w))$ satisfies

$$\hat{X}(\gamma,t,w)\hat{y}(\gamma,t,w) = \gamma^2 u(\gamma,t,w), \tag{5}$$

$$\hat{y}(\gamma,t,w) = M\hat{x}(\gamma,t,w) + \gamma\hat{x}(\gamma,t,w) + q, \tag{6}$$

the next lemma can be easily proved by differentiating (5) and (6) l times with respect to γ and using the fact $u(t,t,w)=w$.

LEMMA 2.5.

$$(\hat{x}+\hat{y})^{(l)}(t,t,w)=\begin{cases} t^2w, & l=0,\\ 2tw+\dfrac{t^2}{1+t}(w-e), & l=1,\\ 2w+\dfrac{4t}{1+t}(w-e), & l=2,\\ \dfrac{6}{1+t}(w-e), & l=3,\\ 0, & l\geq 4, \end{cases}$$

and

$$\hat{y}^{(l)}(t,t,w)=(M+tI)\hat{x}^{(l)}(t,t,w)+l\hat{x}^{(l-1)}(t,t,w),\quad l\geq 1.$$

The following lemma shows how to find $\hat{x}^{(l)}(t,t,w)$ and $\hat{y}^{(l)}(t,t,w)$ assuming we have $(x(t,w),y(t,w),t,w)>0$.

LEMMA 2.6. *$(\hat{x}^{(l)}(t,t,w),\hat{y}^{(l)}(t,t,w))$ is the unique solution to the linear system for $l\geq 1$:*

$$\begin{bmatrix} Y(t,w) & X(t,w)\\ -(M+tI) & I \end{bmatrix}\begin{bmatrix} \hat{x}^{(l)}(t,t,w)\\ \hat{y}^{(l)}(t,t,w)\end{bmatrix}$$

$$=\begin{bmatrix} (\hat{x}+\hat{y})^{(l)}(t,t,w)-\sum_{i=1}^{l-1}\begin{pmatrix} l\\ i\end{pmatrix}\hat{X}^{(i)}(t,t,w)\hat{y}^{(l-i)}(t,t,w)\\ l\hat{x}^{(l-1)}(t,t,w)\end{bmatrix}.$$

Proof. This result simply follows from Lemmas 2.2 and 2.5 and the fact that $(\hat{x}(t,t,w),\hat{y}(t,t,w))=(x(t,w),y(t,w))$. ∎

Therefore each $(\hat{x}^{(l)}(t,t,w),\hat{y}^{(l)}(t,t,w))$ uniquely solves a linear system. And all these linear systems share the same coefficient matrix. Furthermore, the right-hand side of the linear system for solving $(\hat{x}^{(l)}(t,t,w),\hat{y}^{(l)}(t,t,w))$ depends only on $(\hat{x}^{(i)}(t,t,w),\hat{y}^{(i)}(t,t,w))$ for $0\leq i\leq l-1$.

Thus given $(x(t,w),y(t,w),t,w)$, we can sequentially solve L linear systems to find $(\hat{x}^{(1)}(t,t,w),\hat{y}^{(1)}(t,t,w))$, $(\hat{x}^{(2)}(t,t,w),\hat{y}^{(2)}(t,t,w))$, $\ldots$, $(\hat{x}^{(L)}(t,t,w),\hat{y}^{(L)}(t,w))$. Since all these linear systems share the same coefficient matrix, the computational cost is much less than solving L independent linear systems.

3. Algorithm. Let $N(\beta)$ be a neighborhood of the regularized central path $(x(t,e),y(t,e))$ defined as

$$N(\beta):=\{(x,y,t)|x\in R^n_{++},y\in R^n_{++},t\in R_{++},$$
$$\|Xy-t^2e\|_\infty\leq\beta t^2,y=Mx+tx+q\},$$

where $\beta>0$.

ALGORITHM 1.

1. *Select an integer* $L \geq 3$ *and three real numbers* $\alpha \in (0,1)$, $\beta \in (0,1)$, *and* $\theta \in (1, \frac{L}{2})$. *Then find* $(x^0, y^0, t_0) \in N(\beta)$. *Set iteration index* $k = 0$.

2. *At the* k*th iteration, we have* $(x^k, y^k, t_k) \in N(\beta)$. *Let* $w^k = X^k y^k / t_k^2$. *Clearly* $(x^k, y^k) = (x(t_k, w^k), y(t_k, w^k))$.

Then using Lemmas 2.5 *and* 2.6 *we solve* L *linear systems*

$$\begin{bmatrix} Y^k & X^k \\ -(M + t_k I) & I \end{bmatrix} \begin{bmatrix} \hat{x}^{(l)}(t_k, t_k, w^k) \\ \hat{y}^{(l)}(t_k, t_k, w^k) \end{bmatrix}$$

$$= \begin{bmatrix} (\hat{x} + \hat{y})^{(l)}(t_k, t_k, w^k) - \sum_{i=1}^{l-1} \begin{pmatrix} l \\ i \end{pmatrix} \hat{X}^{(i)}(t_k, t_k, w^k) \hat{y}^{(l-i)}(t_k, t_k, w^k) \\ l\hat{x}^{(l-1)}(t_k, t_k, w^k) \end{bmatrix}$$

to find $\left(\hat{x}^{(l)}(t_k, t_k, w^k), \hat{y}^{(l)}(t_k, t_k, w^k)\right)$ *for* $l = 1, 2, \ldots, L$.

Define

$$f^k(s) := \sum_{i=0}^{L} \frac{(s - t_k)^i}{i!} \hat{x}^{(i)}(t_k, t_k, w^k),$$

$$g^k(s) := \sum_{i=0}^{L} \frac{(s - t_k)^i}{i!} \hat{y}^{(i)}(t_k, t_k, w^k) + \frac{(s - t_k)^{L+1}}{L!} \hat{x}^{(L)}(t_k, t_k, w^k).$$

Let $\delta_k = t_k - \min\{(t_k)^\theta, \frac{t_k}{2}\}$. *We pick* t_{k+1} *as the first of* $\{t_k - \alpha^i \delta_k | i = 0, 1, \ldots\}$ *satisfying* $f^k(t_{k+1}) > 0$ *and* $\|F^k(t_{k+1}) g^k(t_{k+1}) - t_{k+1}^2 e\|_\infty \leq \beta {t_{k+1}}^2$.

Then we set $\left(x^{k+1}, y^{k+1}\right) = \left(f^k(t_{k+1}), g^k(t_{k+1})\right)$ *and show that* $\left(x^{k+1}, y^{k+1}, t_{k+1}\right) \in N(\beta)$.

Set $k = k + 1$. *Continue step* 2.

Before we study the convergence properties of this algorithm in next two sections, we first show its feasibility.

3.1. Step 1. There are many ways to find the starting point (x^0, y^0, t_0). For example, we can first find $t_0 > 0$ satisfying

$$\|t_0 Me + \sqrt{t_0} q\|_\infty \leq \beta t_0^2.$$

Then let $x^0 = \sqrt{t_0} e > 0$ and $y^0 = Mx^0 + t_0 x^0 + q$. We have

$$\|X^0 y^0 - t_0^2 e\|_\infty = \|t_0 Me + \sqrt{t_0} q\|_\infty \leq \beta t_0^2.$$

Since $x^0 > 0$ and $\beta \in (0,1)$, we must have $y^0 > 0$. Therefore $(x^0, y^0, t_0) \in N(\beta)$.

3.2. Step 2 . To simplify notation, we omit the super-/subscript k in this subsection; i.e., we use $(x, y, t, w, f, g, \delta)$ to denote $(x^k, y^k, t_k, w^k, f^k, g^k, \delta_k)$. Furthermore, we denote $\left(\hat{x}^{(l)}(t, t, w), \hat{y}^{(l)}(t, t, w), (\hat{x} + \hat{y})^{(l)}(t, t, w)\right)$ by $\left(\hat{x}^{(l)}, \hat{y}^{(l)}, (\hat{x} + \hat{y})^{(l)}\right)$, since (t, w) is fixed in this subsection. We have the following lemmas.

LEMMA 3.1.

$$g(s) = Mf(s) + sf(s) + q.$$

Proof. We have

$$\begin{aligned}
Mf(s)+sf(s)+q &= \sum_{i=0}^{L}\frac{(s-t)^i}{i!}M\hat{x}^{(i)}+\sum_{i=0}^{L}\frac{s(s-t)^i}{i!}\hat{x}^{(i)}+q\\
&= \sum_{i=0}^{L}\frac{(s-t)^i}{i!}M\hat{x}^{(i)}+\sum_{i=0}^{L}\frac{(s-t+t)(s-t)^i}{i!}\hat{x}^{(i)}+q\\
&= \sum_{i=0}^{L}\frac{(s-t)^i}{i!}(M+tI)\hat{x}^{(i)}+\sum_{i=0}^{L}\frac{(s-t)^{i+1}}{i!}\hat{x}^{(i)}+q\\
&= \sum_{i=0}^{L}\frac{(s-t)^i}{i!}(M+tI)\hat{x}^{(i)}+\sum_{i=1}^{L+1}\frac{i(s-t)^i}{i!}\hat{x}^{(i-1)}+q\\
&= (M+tI)\hat{x}^{(0)}+q+\sum_{i=1}^{L}\frac{(s-t)^i}{i!}\left((M+tI)\hat{x}^{(i)}+i\hat{x}^{(i-1)}\right)\\
&\quad+\frac{(s-t)^{L+1}}{L!}\hat{x}^{(L)}\\
&= \hat{y}^{(0)}+\sum_{i=1}^{L}\frac{(s-t)^i}{i!}\hat{y}^{(i)}+\frac{(s-t)^{L+1}}{L!}\hat{x}^{(L)}\\
&= g(s),
\end{aligned}$$

where we use Lemma 2.5. □

LEMMA 3.2.

$$F(s)g(s)=s^2u(s,t,w)+(s-t)^{L+1}R(s,t,w),$$

where

$$\begin{aligned}
R(s,t,w) &= \sum_{l=1}^{L}\sum_{i=0}^{l-1}\frac{(s-t)^i}{l!(i+L-l+1)!}\hat{X}^{(l)}\hat{y}^{(i+L-l+1)}\\
&\quad+\frac{1}{L!}\left(X^{(L)}\sum_{i=0}^{L}\frac{(s-t)^i}{i!}\hat{x}^{(i)}\right).
\end{aligned}\tag{7}$$

Proof. Since

$$\begin{aligned}
&\left(\sum_{i=0}^{L}\frac{(s-t)^i}{i!}\hat{X}^{(i)}\right)\left(\sum_{i=0}^{L}\frac{(s-t)^i}{i!}\hat{y}^{(i)}\right)\\
&\quad=\sum_{l=0}^{L}\sum_{i=0}^{l}\frac{(s-t)^i}{i!}\frac{(s-t)^{l-i}}{(l-i)!}\hat{X}^{(i)}\hat{y}^{(l-i)}+\sum_{l=1}^{L}\sum_{i=L-l+1}^{L}\frac{(s-t)^l}{l!}\frac{(s-t)^i}{i!}\hat{X}^{(l)}\hat{y}^{(i)}\\
&\quad=\sum_{l=0}^{L}\frac{(s-t)^l}{l!}\sum_{i=0}^{l}\begin{pmatrix} l\\ i\end{pmatrix}\hat{X}^{(i)}\hat{y}^{(l-i)}\\
&\qquad+(s-t)^{L+1}\sum_{l=1}^{L}\sum_{i=0}^{l-1}\frac{(s-t)^i}{l!(i+L-l+1)!}\hat{X}^{(l)}\hat{y}^{(i+L-l+1)}\\
&\quad=\sum_{l=0}^{L}\frac{(s-t)^l}{l!}(\hat{x}+\hat{y})^{(l)}+(s-t)^{L+1}\sum_{l=1}^{L}\sum_{i=0}^{l-1}\frac{(s-t)^i}{l!(i+L-l+1)!}\hat{X}^{(l)}\hat{y}^{(i+L-l+1)}\\
&\quad=\sum_{l=0}^{3}\frac{(s-t)^l}{l!}(\hat{x}+\hat{y})^{(l)}+(s-t)^{L+1}\sum_{l=1}^{L}\sum_{i=0}^{l-1}\frac{(s-t)^i}{l!(i+L-l+1)!}\hat{X}^{(l)}\hat{y}^{(i+L-l+1)}\\
&\quad=t^2w+(s-t)\left(2tw+\frac{t^2}{1+t}(w-e)\right)+\frac{(s-t)^2}{2}\left(2w+\frac{4t}{1+t}(w-e)\right)\\
&\qquad+\frac{(s-t)^3}{6}\frac{6}{1+t}(w-e)\\
&\qquad+(s-t)^{L+1}\sum_{l=1}^{L}\sum_{i=0}^{l-1}\frac{(s-t)^i}{l!(i+L-l+1)!}\hat{X}^{(l)}\hat{y}^{(i+L-l+1)}\\
&\quad=s^2u(s,t,w)+(s-t)^{L+1}\sum_{l=1}^{L}\sum_{i=0}^{l-1}\frac{(s-t)^i}{l!(i+L-l+1)!}\hat{X}^{(l)}\hat{y}^{(i+L-l+1)},
\end{aligned}$$

then

$$\begin{aligned}
F(s)g(s)&=\left(\sum_{i=0}^{L}\frac{(s-t)^i}{i!}\hat{X}^{(i)}\right)\left(\sum_{i=0}^{L}\frac{(s-t)^i}{i!}\hat{y}^{(i)}+\frac{(s-t)^{L+1}}{L!}\hat{x}^{(L)}\right)\\
&=\left(\sum_{i=0}^{L}\frac{(s-t)^i}{i!}\hat{X}^{(i)}\right)\left(\sum_{i=0}^{L}\frac{(s-t)^i}{i!}\hat{y}^{(i)}\right)\\
&\qquad+\frac{(s-t)^{L+1}}{L!}\left(X^{(L)}\sum_{i=0}^{L}\frac{(s-t)^i}{i!}\hat{x}^{(i)}\right)\\
&=s^2u(s,t,w)+(s-t)^{L+1}R(s,t,w).\qquad\square
\end{aligned}$$

LEMMA 3.3. *For $0 < s \le t$, we have*

$$\|F(s)g(s) - s^2 e\|_\infty - \beta s^2 \le -(t-s)\left(\frac{\beta s^2}{1+t} - (t-s)^L \|R(s,t,w)\|_\infty\right).$$

Proof. Since $(x, y, t) \in N(\beta)$, then $\|w - e\|_\infty \le \beta$. We have

$$\begin{aligned}
&\|F(s)g(s) - s^2 e\|_\infty - \beta s^2 \\
&\le \|F(s)g(s) - s^2 u(s,t,w)\|_\infty + s^2 \|u(s,t,w) - e\|_\infty - \beta s^2 \\
&= (t-s)^{L+1} \|R(s,t,w)\|_\infty + s^2 \left(1 + \frac{s-t}{1+t}\right) \|w - e\|_\infty - \beta s^2 \\
&\le (t-s)^{L+1} \|R(s,t,w)\|_\infty + s^2 \left(1 + \frac{s-t}{1+t}\right) \beta - \beta s^2 \\
&= -(t-s)\left(\frac{\beta s^2}{1+t} - (t-s)^L \|R(s,t,w)\|_\infty\right). \qquad \square
\end{aligned}$$

Since

$$\lim_{i \longrightarrow \infty} f(t - \alpha^i \delta) = f(t) = \hat{x}^{(0)}(t,t,w) = x > 0,$$

and from Lemma 3.3

$$\limsup_{i \longrightarrow \infty} \frac{\|F(t-\alpha^i\delta)g(t-\alpha^i\delta) - (t-\alpha^i\delta)^2 e\|_\infty - \beta(t-\xi^i\delta)^2}{t - (t - \xi^i \delta)} \le -\frac{\beta t^2}{1+t} < 0,$$

t_{k+1} can be found after a finite number of trials. In addition, we have $x^{k+1} > 0$ and $\|X^{k+1} y^{k+1} - t_{k+1}^2 e\|_\infty \le \beta t_{k+1}^2$. Because $\beta \in (0,1)$, so $y^{k+1} > 0$. Moreover, Lemma 3.1 gives $y^{k+1} = M x^{k+1} + t_{k+1} x^{k+1} + q$. Therefore $(x^{k+1}, y^{k+1}, t_{k+1}) \in N(\beta)$.

4. Global convergence analysis. We need the following assumption for the global and local convergence analysis.

Assumption 1. The LCP (1) has at least one solution.

In this section we show that $\{x^k\}$ converges to the unique least 2-norm solution x^*. First we need to show that t_k decreases to 0.

THEOREM 4.1. *Under Assumption* 1 *we have*

$$\lim_{k \longrightarrow \infty} t_k = 0.$$

Proof. Since $0 < t_{k+1} < t_k$, then $t_k \longrightarrow t_* \ge 0$. Now we prove $t_* = 0$ by contradiction.

Assume $t_* > 0$. Since $(\hat{x}(\gamma,t,w), \hat{y}(\gamma,t,w))$ is an analytic vector function on

$$\left\{(\gamma,t,w) \,\middle|\, w \in R^n_{++}, t \in R_{++}, 0 < \gamma < t + \frac{(1+t)\min\{w_i\}}{1 + \min\{w_i\}}\right\},$$

by the definition of $R(s,t,w)$ we know that $R(s,t,w)$ is bounded on the compact set

$$\left\{(s,t,w) \,\middle|\, \|w - e\|_\infty \le \beta, t_* \le t \le t_0, t - \min\left\{t^\theta, \frac{t}{2}\right\} \le s \le t\right\}.$$

In other words, there exists $C > 0$ such that $\|R(s,t,w)\|_\infty \le C$ when (s,t,w) lies in the above set.

Let k be sufficiently large such that t_k satisfies

$$0 < \epsilon := \left(\frac{\min\{t_*^\theta, \frac{t_*}{2}\}^2\beta}{2C(1+t_*)}\right)^{\frac{1}{L}} < \left(\frac{\min\{t_k^\theta, \frac{t_k}{2}\}^2\beta}{C(1+t_k)}\right)^{\frac{1}{L}}$$

and

$$\min\left\{t_k^\theta, \frac{t_k}{2}\right\} < t_* < t_k < t_* + \frac{\alpha}{2}\epsilon.$$

Using Lemma 3.3, for all $s \in [\min\{t_k^\theta, \frac{t_k}{2}\}, t_k] \cap [t_k - \epsilon, t_k]$ we have

$$\begin{aligned}\|F^k(s)g^k(s) - s^2 e\|_\infty &\le \beta s^2 - (t_k - s)\left(\frac{\beta s^2}{1+t_k} - C(t_k - s)^L\right)\\ &\le \beta s^2 - (t_k - s)\left(\frac{\beta \min\{t_k^\theta, \frac{t_k}{2}\}^2}{1+t_k} - C\epsilon^L\right)\\ &\le \beta s^2.\end{aligned}$$

Then from the continuity of $(f^k(s), g^k(s))$ and the fact that $(f^k(t_k), g^k(t_k)) = (x^k, y^k) > 0$, we know that $(f^k(s), g^k(s)) > 0$ for all $s \in [\min\{t_k^\theta, \frac{t_k}{2}\}, t_k] \cap [t_k - \epsilon, t_k]$.

If $\epsilon > \delta_k = t_k - \min\{t_k^\theta, \frac{t_k}{2}\}$, then $t_{k+1} = \min\{t_k^\theta, \frac{t_k}{2}\} < t_*$, which is a contradiction. So $\epsilon \le \delta_k$.

Let i be the integer such that $t_{k+1} = t_k - \alpha^i\delta_k$. Since $t_k - \alpha^0\delta_k = \min\{{t_k}^\theta, \frac{t_k}{2}\} < t_* < t_{k+1}$, then $i \ge 1$. By the definition of t_{k+1}, we must have $t_k - \alpha^{i-1}\delta_k < t_k - \epsilon$, and hence $\epsilon < \alpha^{i-1}\delta_k$.

Therefore we have

$$\begin{aligned}t_{k+1} = t_k - \alpha^i\delta^k > t_* &\Longrightarrow \frac{\alpha}{2}\epsilon > t_k - t_* > \alpha^i\delta^k > \alpha\epsilon\\ &\Longrightarrow \frac{1}{2} > 1\\ &\Longrightarrow \text{contradiction.}\end{aligned}$$

So $t_* = 0$. ∎

Now we can show the global convergence of $\{x^k\}$.

THEOREM 4.2. *Under Assumption* 1 *we have*

$$\lim_{k\longrightarrow\infty} x^k = x^*,$$

where x^* is the unique least 2*-norm solution of the monotone LCP* (1).

Proof. Since $(x^k, y^k, t_k) \in N(\beta)$, then $(1-\beta){t_k}^2 e \le X^k y^k \le (1+\beta){t_k}^2 e$, and hence ${x^k}^T y^k \le (1+\beta)n{t_k}^2$.

Using the facts that $0 \leq x^* \perp y^* := Mx^* + q \geq 0$, $x^k > 0$, and $y^k = Mx^k + t_k x^k + q > 0$, we have

$$\begin{aligned}(1+\beta)nt_k{}^2 &\geq x^{k^T} y^k - x^{k^T} y^* - x^{*^T} y^k + x^{*^T} y^* \\ &= (x^k - x^*)^T (y^k - y^*) \\ &= (x^k - x^*)^T \left(M(x^k - x^*) + t_k x^k\right) \\ &\geq t_k (x^k - x^*)^T x^k.\end{aligned}$$

Hence

$$(1+\beta)nt_k \geq \|x^k\|^2 - \|x^*\|\|x^k\|. \tag{8}$$

Since $0 < t_k < t_0$, then $\{x^k\}$ is bounded. Let $\bar{x}$ be any limit point of $\{x^k\}$. Because $t_k \longrightarrow 0$, it is easy to see that $\bar{x}$ is a solution to the LCP. Hence $\|\bar{x}\| \geq \|x^*\|$. From (8) we can get

$$0 \geq \|\bar{x}\|^2 - \|x^*\|\|\bar{x}\| = \|\bar{x}\|(\|\bar{x}\| - \|x^*\|).$$

So $\|\bar{x}\| = \|x^*\|$. By the uniqueness of the least 2-norm solution, we must have $\bar{x} = x^*$. Since $\bar{x}$ is any limit point of $\{x^k\}$, we then have

$$\lim_{k \longrightarrow \infty} x^k = x^*. \qquad \square$$

5. Local convergence analysis. First we introduce the following notation for this section:

(i) $S :=$ the solution set of the LCP (1).
(ii) $S_y := \{y \in R^n | y = Mx + q$ for some $x \in \mathcal{S}\}$.
(iii) $B := \{i | x_i > 0$ for some $x \in S\}$.
(iv) $N := \{i | y_i > 0$ for some $y \in S_y\}$.
(v) $J := \{i | x_i = y_i = 0 \ \forall x \in S, \forall y \in S_y\}$.

It is well known that the sets B, N, J form a partition of the index set $\{1, 2, \ldots, n\}$; i.e., they are pairwise disjoint and $B \cup N \cup J = \{1, 2, \ldots, n\}$. Therefore a solution x is called a maximally complementary solution if

$$x_B > 0 \quad \text{and} \quad (Mx + q)_N > 0.$$

If $J = \emptyset$, then a maximally complementary solution is called a strictly complementary solution. Note that if $J \neq \emptyset$, then LCP (1) has no strictly complementary solution.

In [24], a superlinearly convergent algorithm based on the same class of paths was proposed under the assumption that x^* is the only solution of the LCP and is strictly complementary. The assumption needed for our local analysis is weaker.

Assumption 2. x^* is maximally complementary.

Our analysis relies on the boundedness of $\{(\hat{x}^{(l)}, \hat{y}^{(l)})(t, t, w) | l = 1, 2, \ldots, L\}$ on the noncompact set $\{(t, w) | 0 < t \leq t_0, \|w - e\|_\infty \leq \beta\}$. But we will show a stronger result that $(\hat{x}(\gamma, t, w), \hat{y}(\gamma, t, w))$, an analytic vector function on

$$\left\{ (\gamma, t, w) \,\middle|\, w > 0, t > 0, 0 < \gamma < t + \frac{(1+t)\min\{w_i\}}{1 + \min\{w_i\}} \right\},$$

can be analytically extended to the set

$$\left\{(\gamma,t,w)\,\middle|\,w>0,t\geq 0,0\leq\gamma<t+\frac{(1+t)\min\{w_i\}}{1+\min\{w_i\}}\right\}.$$

From Theorem 2.1, for any fixed $w>0$ we have

$$\lim_{t\longrightarrow 0+}(x(t,w),y(t,w))=(x^*,y^*).$$

Now we first show that $(x(t,w),y(t,w))$ can be analytically extended to $\{(t,w)|t\geq 0,w>0\}$. The following two technical lemmas will be needed.

LEMMA 5.1. *For any fixed $w>0$, we have $x_j(t,w)=\Theta(t)$ and $y_j(t,w)=\Theta(t)$ for all $j\in J$ as $t\longrightarrow 0+$. Here we use the standard big-O notation: Given functions $f_1(t)$ and $f_2(t)$, we say $f_1(t)=O(t)$ and $f_2(t)=\Theta(t)$ as $t\longrightarrow 0+$ if and only if there exist positive constants C_1, C_2, and C_3 such that when t is sufficiently small we have $|f_1(t)|\leq C_1t$, and $C_2t\leq|f_2(t)|\leq C_3t$.*

Proof. For any $x,y\in R^n$, let $d(x,S)$ and $d(y,S_y)$ denote the distance between x and S and the distance between y and S_y, respectively.

Corollary 2.2 of [8] states that $r(x)+s(x)$ is a global error bound for a monotone LCP, where $r(x)=\|x-(x-Mx-q)_+\|$ and $s(x)=\|(-Mx-q,-x,X(Mx+q))_+\|$. So there exists $\tau>0$ such that for any $x\in R^n$, $d(x,S)\leq\tau(r(x)+s(x))$.

For fixed $w>0$, we consider $(x(t,w),y(t,w))>0$, the solution of

$$\begin{cases}X(t,w)y(t,w)=t^2w,\\ y(t,w)=Mx(t,w)+tx(t,w)+q,\end{cases}$$

where $0<t\leq 1$.

Since $(x(t,w),y(t,w))\longrightarrow(x^*,y^*)$ as $t\longrightarrow 0+$, and $(x(t,w),y(t,w))$ is an analytic vector function on $t>0$, then $\{(x(t,w),y(t,w))|0<t\leq 1\}$ is bounded.

Now we show that $s(x(t,w))=O(t)$ and $r(x(t,w))=O(t)$.

For $s(x(t,w))$ we have

$$\begin{aligned}(X(t,w)(Mx(t,w)+q))_+&=(X(t,w)(y(t,w)-tx(t,w)))_+\\&=(t^2w-tX(t,w)^2e)_+\leq t^2w=O(t^2),\\(-x(t,w))_+&=0,\\(-Mx(t,w)-q)_+&=(tx(t,w)-y(t,w))_+\leq tx(t,w)=O(t),\end{aligned}$$

and so $s(x(t,w))=O(t)$.

For $r(x(t,w))$ we have $r(x(t,w))=\|x(t,w)-((1+t)x(t,w)-y(t,w))_+\|$.

(i) If i is an index such that $(1+t)x_i(t,w)-y_i(t,w)>0$, then $0<y_i(t,w)<(1+t)x_i(t,w)\leq 2x_i(t,w)$, and so $y_i(t,w)^2<2x_i(t,w)y_i(t,w)=2w_it^2$, and hence $y_i(t,w)\leq\sqrt{2w_i}t$; then

$$\begin{aligned}|x_i(t,w)-((1+t)x_i(t,w)-y_i(t,w))_+|&=|y_i(t,w)-tx_i(t,w)|\\&\leq y_i(t,w)+tx_i(t,w)\\&\leq(\sqrt{2w_i}+\|x(t,w)\|)t.\end{aligned}$$

(ii) If i is an index such that $(1+t)x_i(t,w)-y_i(t,w) \le 0$, then $x_i(t,w) \le y_i(t,w)$, $x_i(t,w)^2 \le x_i(t,w)y_i(t,w) = w_i t^2$, and so $x_i(t,w) \le \sqrt{w_i}t$, and hence

$$|x_i(t,w) - ((1+t)x_i(t,w) - y_i(t,w))_+| = x_i(t,w) \le \sqrt{w_i}t.$$

Combining the above inequalities we have $r(x(t,w)) = O(t)$. Therefore $d(x(t,w), S) = O(t)$.

On the other hand, we also have

$$\begin{aligned} d(y(t,w), S_y) &= \min\{\|Mx(t,w) + tx(t,w) + q - (Mx+q)\| \,|x \in S\} \\ &= \min\{\|M(x(t,w)-x) + tx(t,w)\| | x \in S\} \\ &\le \min\{\|M\|\|x(t,w)-x\| + t\|x(t,w)\| | x \in S\} \\ &\le \|M\| d(x(t,w), S) + t\|x(t,w)\| \\ &= O(t). \end{aligned}$$

For all $j \in J$, since $x_j = y_j = 0$ for all $x \in S$ and $y \in S_y$, we have

$$0 < x_j(t,w) \le d(x(t,w), S) = O(t),$$

$$0 < y_j(t,w) \le d(y(t,w), S_y) = O(t).$$

Notice that $x_j(t,w)y_j(t,w) = t^2 w_j$ for all $t > 0$, we must have $x_j(t,w) = \Theta(t)$, and $y_j(t,w) = \Theta(t)$. □

LEMMA 5.2. *If $0 \preceq M \in R^{n\times n}$, then $\mathcal{N}(M) \cap \mathcal{R}(M) = \{0\}$, where $\mathcal{N}(M)$ and $\mathcal{R}(M)$ represent the null space and range space of M, respectively.*

Proof. Since $\mathcal{N}(M) \subseteq \mathcal{N}(M^2)$, $\dim(\mathcal{N}(M)) = n - \operatorname{rank}(M)$, and $\dim(\mathcal{N}(M^2)) = n - \operatorname{rank}(M^2)$, we have

$$\mathcal{N}(M) \cap \mathcal{R}(M) = \{0\} \Longleftrightarrow \mathcal{N}(M) = \mathcal{N}(M^2) \Longleftrightarrow \operatorname{rank}(M) = \operatorname{rank}(M^2).$$

Let $A = \frac{M+M^T}{2}$, $B = \frac{M-M^T}{2}$. Then A is symmetric positive semidefinite, while B is skew-symmetric. If A is nonsingular, then so is M. Hence $\mathcal{N}(M) = \mathcal{N}(M^2) = \{0\}$. So we need only to consider the case when A is singular.

If A is of the form $\left[\begin{smallmatrix} D & 0 \\ 0 & 0 \end{smallmatrix}\right]$, where D is symmetric positive definite, then we can partition B likewise: $B = \left[\begin{smallmatrix} C & F \\ -F^T & G \end{smallmatrix}\right]$, where both C and G are skew-symmetric.

Letting $z = \left[\begin{smallmatrix} u \\ v \end{smallmatrix}\right] \in \mathcal{N}(M)$, we have

$$Mz = \begin{bmatrix} D+C & F \\ -F^T & G \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} (D+C)u + Fv \\ -F^T u + Gv \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

So $u^T((D+C)u + Fv) + v^T(-F^T u + Gv) = u^T D u = 0$, and hence we get $u = 0$, $Fv = 0$, and $Gv = 0$.

Therefore

$$\mathcal{N}(M) = \left\{ \begin{bmatrix} 0 \\ v \end{bmatrix} \middle| \, v \in \mathcal{N}(F) \cap N(G) \right\}.$$

Now if $z = \left[\begin{smallmatrix} u \\ v \end{smallmatrix}\right] \in \mathcal{N}(M^2)$, then $Mz = \left[\begin{smallmatrix} (D+C)u + Fv \\ -F^T u + Gv \end{smallmatrix}\right] \in \mathcal{N}(M)$.

So $(D+C)u+Fv=0$, $F(-F^Tu+Gv)=0$, and $G(-F^Tu+Gv)=0$. Then we have

$$\begin{aligned}\|-F^Tu+Gv\|^2 &= (-u^TF+v^TG^T)(-F^Tu+Gv)\\ &= -u^TF(-F^Tu+Gv)+v^TG^T(-F^Tu+Gv)\\ &= -u^TF(-F^Tu+Gv)-v^TG(-F^Tu+Gv)\\ &= 0.\end{aligned}$$

Thus $-F^Tu+Gv=0$, and so $Mz=0$, $z\in\mathcal{N}(M)$. Therefore $\mathcal{N}(M)=\mathcal{N}(M^2)$, which is equivalent to $\operatorname{rank}(M)=\operatorname{rank}(M^2)$ when the symmetric part of M has this special form.

For general symmetric positive semidefinite matrix A, there exists an orthogonal matrix O such that $O^TAO=\left[\begin{smallmatrix} D & 0\\ 0 & 0\end{smallmatrix}\right]$, where D is symmetric positive definite. So $\operatorname{rank}(O^TMO)=\operatorname{rank}((O^TMO)^2)$.

Thus $\operatorname{rank}(M)=\operatorname{rank}(O^TMO)=\operatorname{rank}((O^TMO)^2)=\operatorname{rank}(O^TM^2O)=\operatorname{rank}(M^2)$, and hence $\mathcal{N}(M)=\mathcal{N}(M^2)$ and $\mathcal{N}(M)\cap\mathcal{R}(M)=\{0\}$. □

Now we can prove the main theorem.

THEOREM 5.3. *If x^* is maximally complementary, i.e., $(x_B^*,y_N^*)>0$, then $(x(t,w),y(t,w))$ can be analytically extended to an open set $\mathcal{P}$ containing $\{t\ge 0,w>0\}$. In other words, there exist analytic vector functions $f(t,w)$ and $g(t,w)$ on $\mathcal{P}$ such that $f(t,w)=x(t,w)$ and $g(t,w)=y(t,w)$ for all $(t,w)>0$.*

Proof. Without loss of generality, we assume $B=\{1,2,\dots,|B|\}$, $J=\{|B|+1,\dots,|B|+|J|\}$, and $N=\{|B|+|J|+1,\dots,n\}$. Then $x_B^*>0$, $y_B^*=0$, $x_N^*=0$, $y_N^*>0$, and $x_J^*=y_J^*=0$.

We consider only $t\le 1$.

Define the following vector functions for $t>0,w>0$:

$$\begin{aligned}\tilde{x}_B(t,w) &:= x_B(t,w),\\ \tilde{x}_N(t,w) &:= W_N y_N(t,w)^{-1}=t^{-2}x_N(t,w),\\ \tilde{x}_J(t,w) &:= t^{-1}x_J(t,w),\\ \tilde{y}_B(t,w) &:= W_B x_B(t,w)^{-1}=t^{-2}y_B(t,w),\\ \tilde{y}_J(t,w) &:= t^{-1}y_J(t,w),\\ \tilde{y}_N(t,w) &:= y_N(t,w),\\ \tilde{x}(t,w) &:= (\tilde{x}_B(t,w),\tilde{x}_J(t,w),\tilde{x}_N(t,w)),\\ \tilde{y}(t,w) &:= (\tilde{y}_B(t,w),\tilde{y}_J(t,w),\tilde{y}_N(t,w)).\end{aligned}$$

These are all positive analytic vector functions on $(t>0,w>0)$.

For fixed w, since $(x_B(t,w),y_N(t,w))\longrightarrow(x_B^*,y_N^*)>0$, and$(x_J(t,w),y_J(t,w))=\Theta(t)$, then $(\tilde{x}(t,w),\tilde{y}(t,w))$ is bounded when $0<t\le 1$. Let $\{t_k(w)\}$ be a positive sequence decreasing to 0 such that $(\tilde{x}(t_k(w),w),\tilde{y}(t_k(w),w))$ is convergent, and define $(\tilde{x}(0,w),\tilde{y}(0,w))$ to be the limit. Since $\tilde{X}(t,w)\tilde{y}(t,w)=w$ and $(\tilde{x}(t,w),\tilde{y}(t,w))>0$ for $t>0$, we must have $(\tilde{x}(0,w),\tilde{y}(0,w))>0$.

Therefore $(\tilde{x}(t,w),\tilde{y}(t,w))$ is the unique positive solution of the following system for any fixed $(t,w)>0$:

$$\begin{cases} \begin{bmatrix} M_{BB} & M_{BJ} & M_{BN} \\ M_{JB} & M_{JJ} & M_{JN} \\ M_{NB} & M_{NJ} & M_{NN} \end{bmatrix} \begin{bmatrix} \tilde{x}_B \\ t\tilde{x}_J \\ t^2\tilde{x}_N \end{bmatrix} + \begin{bmatrix} t\tilde{x}_B \\ t^2\tilde{x}_J \\ t^3\tilde{x}_N \end{bmatrix} - \begin{bmatrix} t^2\tilde{y}_B \\ t\tilde{y}_J \\ \tilde{y}_N \end{bmatrix} + \begin{bmatrix} q_B \\ q_J \\ q_N \end{bmatrix} = 0, \\ \tilde{X}\tilde{y} - w = 0, \end{cases}$$

while $(\tilde{x}(0,w),\tilde{y}(0,w))$ is a solution of this system for fixed $(t=0,w>0)$.

The Jacobian of this system with respect to $(\tilde{x}_B,\tilde{x}_J,\tilde{x}_N,\tilde{y}_B,\tilde{y}_J,\tilde{y}_N)$ is

$$\begin{bmatrix} M_{BB}+tI_B & tM_{BJ} & t^2M_{BN} & -t^2I_B & 0 & 0 \\ M_{JB} & tM_{JJ}+t^2I_J & t^2M_{JN} & 0 & -tI_J & 0 \\ M_{NB} & tM_{NJ} & t^2M_{NN}+t^3I_N & 0 & 0 & -I_N \\ \tilde{Y}_B(t,w) & 0 & 0 & \tilde{X}_B(t,w) & 0 & 0 \\ 0 & \tilde{Y}_J(t,w) & 0 & 0 & \tilde{X}_J(t,w) & 0 \\ 0 & 0 & \tilde{Y}_N(t,w) & 0 & 0 & \tilde{X}_N(t,w) \end{bmatrix}.$$

This matrix may become singular when $t=0$, which prevents us from using the implicit function theorem as in Theorem 2.3. In order to fix this problem, we use the technique of "adding redundant equations" which was used in [3, 2, 15, 13] to study the analyticity of the central path of LP, SDP, LCP, and SDLCP.

Let $r=\operatorname{rank}(\left[\begin{smallmatrix} M_{BB} \\ M_{JB} \end{smallmatrix}\right])$. Then there exists $P\in R^{(|B|+|J|-r)\times(|B|+|J|)}$ such that $P\left[\begin{smallmatrix} M_{BB} \\ M_{JB} \end{smallmatrix}\right]=0$, and $\mathcal{N}(P)=\mathcal{R}\big(\left[\begin{smallmatrix} M_{BB} \\ M_{JB} \end{smallmatrix}\right]\big)$.

Multiplying the first two equations by P from the left, we get

$$P\begin{bmatrix} M_{BB} & M_{BJ} & M_{BN} \\ M_{JB} & M_{JJ} & M_{JN} \end{bmatrix}\begin{bmatrix} \tilde{x}_B \\ t\tilde{x}_J \\ t^2\tilde{x}_N \end{bmatrix} + P\begin{bmatrix} t\tilde{x}_B \\ t^2\tilde{x}_J \end{bmatrix}$$

$$-P\begin{bmatrix} t^2\tilde{y}_B \\ t\tilde{y}_J \end{bmatrix} + P\begin{bmatrix} q_B \\ q_J \end{bmatrix} = 0$$

$\Longrightarrow$

$$P\begin{bmatrix} M_{BJ} & M_{BN} \\ M_{JJ} & M_{JN} \end{bmatrix}\begin{bmatrix} t\tilde{x}_J \\ t^2\tilde{x}_N \end{bmatrix} + P\begin{bmatrix} t\tilde{x}_B \\ t^2\tilde{x}_J \end{bmatrix} - P\begin{bmatrix} t^2\tilde{y}_B \\ t\tilde{y}_J \end{bmatrix} + P\begin{bmatrix} q_B \\ q_J \end{bmatrix} = 0.$$

Letting $t\longrightarrow 0$, we get $P\left[\begin{smallmatrix} q_B \\ q_J \end{smallmatrix}\right]=0$. Hence the above system is equivalent to

$$P\begin{bmatrix} M_{BJ} & M_{BN} \\ M_{JJ} & M_{JN} \end{bmatrix}\begin{bmatrix} t\tilde{x}_J \\ t^2\tilde{x}_N \end{bmatrix} + P\begin{bmatrix} t\tilde{x}_B \\ t^2\tilde{x}_J \end{bmatrix} - P\begin{bmatrix} t^2\tilde{y}_B \\ t\tilde{y}_J \end{bmatrix} = 0.$$

Factoring out t, we get

$$P\begin{bmatrix} M_{BJ} & M_{BN} \\ M_{JJ} & M_{JN} \end{bmatrix}\begin{bmatrix} \tilde{x}_J \\ t\tilde{x}_N \end{bmatrix} + P\begin{bmatrix} \tilde{x}_B \\ t\tilde{x}_J \end{bmatrix} - P\begin{bmatrix} t\tilde{y}_B \\ \tilde{y}_J \end{bmatrix} = 0.$$

Therefore $(\tilde{x}(t,w), \tilde{y}(t,w), t, w)$ solves the enlarged system for $t \geq 0$, $w > 0$:

$$\Phi(\tilde{x}, \tilde{y}, t, w) := \begin{pmatrix} P\begin{bmatrix} M_{BJ} & M_{BN} \\ M_{JJ} & M_{JN} \end{bmatrix}\begin{bmatrix} \tilde{x}_J \\ t\tilde{x}_N \end{bmatrix} + P\begin{bmatrix} \tilde{x}_B \\ t\tilde{x}_J \end{bmatrix} - P\begin{bmatrix} t\tilde{y}_B \\ \tilde{y}_J \end{bmatrix} \\ \begin{bmatrix} M_{BB} & M_{BJ} & M_{BN} \\ M_{JB} & M_{JJ} & M_{JN} \\ M_{NB} & M_{NJ} & M_{NN} \end{bmatrix}\begin{bmatrix} \tilde{x}_B \\ t\tilde{x}_J \\ t^2\tilde{x}_N \end{bmatrix} + \begin{bmatrix} t\tilde{x}_B \\ t^2\tilde{x}_J \\ t^3\tilde{x}_N \end{bmatrix} - \begin{bmatrix} t^2\tilde{y}_B \\ t\tilde{y}_J \\ \tilde{y}_N \end{bmatrix} + \begin{bmatrix} q_B \\ q_J \\ q_N \end{bmatrix} \\ \tilde{X}\tilde{y} - w \end{pmatrix} = 0.$$

We partition P as $P = [P_B, P_J]$; then the Jacobian of $\Phi(\tilde{x}, \tilde{y}, t, w)$ with respect to $(\tilde{x}_B, \tilde{x}_J, \tilde{x}_N, \tilde{y}_B, \tilde{y}_J, \tilde{y}_N)$ is

$$\Phi'(\tilde{x}_B, \tilde{x}_J, \tilde{x}_N, \tilde{y}_B, \tilde{y}_J, \tilde{y}_N, t, w)$$

$$= \begin{bmatrix} P_B & P\begin{bmatrix} M_{BJ} \\ M_{JJ} \end{bmatrix} + tP_J & tP\begin{bmatrix} M_{BN} \\ M_{JN} \end{bmatrix} & -tP_B & -P_J & 0 \\ M_{BB} + tI_B & tM_{BJ} & t^2M_{BN} & -t^2I_B & 0 & 0 \\ M_{JB} & tM_{JJ} + t^2I_J & t^2M_{JN} & 0 & -tI_J & 0 \\ M_{NB} & tM_{NJ} & t^2M_{NN} + t^3I_N & 0 & 0 & -I_N \\ \tilde{Y}_B(t,w) & 0 & 0 & \tilde{X}_B(t,w) & 0 & 0 \\ 0 & \tilde{Y}_J(t,w) & 0 & 0 & \tilde{X}_J(t,w) & 0 \\ 0 & 0 & \tilde{Y}_N(t,w) & 0 & 0 & \tilde{X}_N(t,w) \end{bmatrix},$$

which is simplified to

$$J = \begin{bmatrix} P_B & P\begin{bmatrix} M_{BJ} \\ M_{JJ} \end{bmatrix} & 0 & 0 & -P_J & 0 \\ M_{BB} & 0 & 0 & 0 & 0 & 0 \\ M_{JB} & 0 & 0 & 0 & 0 & 0 \\ M_{NB} & 0 & 0 & 0 & 0 & -I_N \\ \tilde{Y}_B(0,w) & 0 & 0 & \tilde{X}_B(0,w) & 0 & 0 \\ 0 & \tilde{Y}_J(0,w) & 0 & 0 & \tilde{X}_J(0,w) & 0 \\ 0 & 0 & \tilde{Y}_N(0,w) & 0 & 0 & \tilde{X}_N(0,w) \end{bmatrix}$$

at $(\tilde{x}(0,w), \tilde{y}(0,w), 0, w)$.

Considering the rank of J (after some simple row/column operations), we have

$$\operatorname{rank}(J) = \operatorname{rank}\left(\begin{bmatrix} P_B & P\begin{bmatrix} M_{BJ} \\ M_{JJ} \end{bmatrix} & 0 & 0 & -P_J & 0 \\ M_{BB} & 0 & 0 & 0 & 0 & 0 \\ M_{JB} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -I_N \\ 0 & 0 & 0 & \tilde{X}_B(0,w) & 0 & 0 \\ 0 & \tilde{Y}_J(0,w) & 0 & 0 & \tilde{X}_J(0,w) & 0 \\ 0 & 0 & \tilde{Y}_N(0,w) & 0 & 0 & 0 \end{bmatrix}\right)$$

$$= 2|N| + |B| + \operatorname{rank}\left(\begin{bmatrix} P_B & P\begin{bmatrix} M_{BJ} \\ M_{JJ} \end{bmatrix} & -P_J \\ M_{BB} & 0 & 0 \\ M_{JB} & 0 & 0 \\ 0 & \tilde{Y}_J(0,w) & \tilde{X}_J(0,w) \end{bmatrix}\right)$$

$$= 2|N| + |B| + \operatorname{rank}\left(\begin{bmatrix} P_B & P\begin{bmatrix} M_{BJ} \\ M_{JJ} \end{bmatrix} + P_J\tilde{X}_J(0,w)^{-1}\tilde{Y}_J(0,w) & 0 \\ M_{BB} & 0 & 0 \\ M_{JB} & 0 & 0 \\ 0 & \tilde{Y}_J(0,w) & \tilde{X}_J(0,w) \end{bmatrix}\right)$$

$$= |N| + n + \operatorname{rank}\left(\begin{bmatrix} P_B & P\begin{bmatrix} M_{BJ} \\ M_{JJ} \end{bmatrix} + P_J\tilde{X}_J(0,w)^{-1}\tilde{Y}_J(0,w) \\ M_{BB} & 0 \\ M_{JB} & 0 \end{bmatrix}\right).$$

Let $Z = \tilde{X}_J(0,w)^{-1}\tilde{Y}_J(0,w)$ and

$$A = \begin{bmatrix} P_B & P\begin{bmatrix} M_{BJ} \\ M_{JJ} \end{bmatrix} + P_JZ \\ M_{BB} & 0 \\ M_{JB} & 0 \end{bmatrix} \in R^{(2|B|+|J|)\times(|B|+|J|)}.$$

Now we show that A has full column rank. For all $\begin{bmatrix} u \\ v \end{bmatrix} \in \mathcal{N}(A)$, we have

$$P_Bu + P\begin{bmatrix} M_{BJ} \\ M_{JJ} \end{bmatrix} v + P_JZv = 0,$$

$$M_{BB}u = 0,$$

$$M_{JB}u = 0.$$

The first equation leads to

$$P\begin{bmatrix} u+M_{BJ}v \\ (M_{JJ}+Z)v \end{bmatrix} = 0 \Longrightarrow \begin{bmatrix} u+M_{BJ}v \\ (M_{JJ}+Z)v \end{bmatrix} \in \mathcal{R}\left(\begin{bmatrix} M_{BB} \\ M_{JB} \end{bmatrix}\right).$$

So there exists $a \in R^{|B|}$ such that $u + M_{BJ}v = M_{BB}a$ and $(M_{JJ} + Z)v = M_{JB}a$.

Since $M \succeq 0$ and $Z \succ 0$, then $M_{JJ} + Z \succ 0$ and $\left[\begin{smallmatrix} M_{BB} & M_{BJ} \\ M_{JB} & M_{JJ}+Z \end{smallmatrix}\right] \succeq 0$. So we have $v = (M_{JJ} + Z)^{-1}M_{JB}a$ and $u = (M_{BB} - M_{BJ}(M_{JJ} + Z)^{-1}M_{JB})a$.

Since

$$\begin{bmatrix} M_{BB} - M_{BJ}(M_{JJ}+Z)^{-1}M_{JB} & M_{BJ} - M_{JB}^T(M_{JJ}+Z)^{-T}(M_{JJ}+Z) \\ 0 & M_{JJ}+Z \end{bmatrix}$$

$$= \begin{bmatrix} I_B & -M_{JB}^T(M_{JJ}+Z)^{-T} \\ 0 & I_J \end{bmatrix} \begin{bmatrix} M_{BB} & M_{BJ} \\ M_{JB} & M_{JJ}+Z \end{bmatrix}$$

$$\times \begin{bmatrix} I_B & 0 \\ -(M_{JJ}+Z)^{-1}M_{JB} & I_J \end{bmatrix}$$

$$\succeq 0,$$

then $M_1 := M_{BB} - M_{BJ}(M_{JJ}+Z)^{-1}M_{JB} \succeq 0$. Now because $u \in \mathcal{N}(M_{BB}) \cap \mathcal{N}(M_{JB})$ and $u \in \mathcal{R}(M_1)$, we have $u \in \mathcal{N}(M_1) \cap \mathcal{R}(M_1)$, and hence by Lemma 5.2, $u = 0$.

So $(M_{BB} - M_{BJ}(M_{JJ} + Z)^{-1}M_{JB})a = 0$, and hence

$$0 = \begin{bmatrix} a^T & 0 \end{bmatrix}$$

$$\times \begin{bmatrix} M_{BB} - M_{BJ}(M_{JJ}+Z)^{-1}M_{JB} & M_{BJ} - M_{JB}^T(M_{JJ}+Z)^{-T}(M_{JJ}+Z) \\ 0 & M_{JJ}+Z \end{bmatrix}$$

$$\times \begin{bmatrix} a \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} a^T & 0 \end{bmatrix} \begin{bmatrix} I_B & -M_{JB}^T(M_{JJ}+Z)^{-T} \\ 0 & I_J \end{bmatrix} \begin{bmatrix} M_{BB} & M_{BJ} \\ M_{JB} & M_{JJ}+Z \end{bmatrix}$$

$$\times \begin{bmatrix} I_B & 0 \\ -(M_{JJ}+Z)^{-1}M_{JB} & I_J \end{bmatrix} \begin{bmatrix} a \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} a^T & -a^T M_{JB}^T(M_{JJ}+Z)^{-T} \end{bmatrix} \begin{bmatrix} M_{BB} & M_{BJ} \\ M_{JB} & M_{JJ}+Z \end{bmatrix}$$

$$\times \begin{bmatrix} a \\ -(M_{JJ}+Z)^{-1}M_{JB}a \end{bmatrix}$$

$$= \begin{bmatrix} a^T & -v^T \end{bmatrix} \left(\begin{bmatrix} M_{BB} & M_{BJ} \\ M_{JB} & M_{JJ} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & Z \end{bmatrix} \right) \begin{bmatrix} a \\ -v \end{bmatrix}$$

$$\geq v^T Z v$$

$$\geq 0.$$

Because $Z \succ 0$, we must have $v = 0$.

Therefore $\mathcal{N}(A) = \{0\}$, and so $\operatorname{rank}(A) = |B| + |J|$, and hence $\operatorname{rank}(J) = 2n$.

Thus the system of $2n+|B|+|J|-r$ equations $\Phi(\tilde{x}, \tilde{y}, t, w) = 0$ contains a subsystem $\Phi_{\mathcal{L}}(\tilde{x}, \tilde{y}, t, w) = 0$ of $|\mathcal{L}| = 2n$ equations such that the Jacobian of the subsystem $\Phi_{\mathcal{L}}(\tilde{x}, \tilde{y}, t, w)$ with respect to $(\tilde{x}, \tilde{y})$ is nonsingular at the solution $(\tilde{x}(0, w), \tilde{y}(0, w), 0, w)$ for any $w > 0$. So by the implicit function theorem, for any fixed $\bar{w} > 0$, the subsystem $\Phi_{\mathcal{L}}(\tilde{x}, \tilde{y}, t, w)$ (depending analytically on every variable) has a locally unique analytic solution $(x_{\bar{w}}(t, w), y_{\bar{w}}(t, w))$. More specifically, there exist open balls $B_{xy}(\bar{w}) \subset R^{2n}$ and $B_{tw}(\bar{w}) \subset R^{n+1}$, centering at $(\tilde{x}(0, \bar{w}), \tilde{y}(0, \bar{w}))$ and $(t = 0, \bar{w})$, respectively, satisfying the following:

(i) For each $(t, w) \in B_{tw}(\bar{w})$, there is a unique solution $(x_{\bar{w}}(t, w), y_{\bar{w}}(t, w)) \in B_{xy}(\bar{w})$ of the system $\Phi_{\mathcal{L}}(\tilde{x}, \tilde{y}, t, w) = 0$.

(ii) $(x_{\bar{w}}(t, w), y_{\bar{w}}(t, w))$ is an analytic vector function for $(t, w) \in B_{tw}(\bar{w})$.

(iii) $(x_{\bar{w}}(0, \bar{w}), y_{\bar{w}}(0, \bar{w})) = (\tilde{x}(0, \bar{w}), \tilde{y}(0, \bar{w}))$.

Since $t_k(\bar{w}) \longrightarrow 0$ and $(\tilde{x}(t_k(\bar{w}), \bar{w}), \tilde{y}(t_k(\bar{w}), \bar{w})) \longrightarrow (\tilde{x}(0, \bar{w}), \tilde{y}(0, \bar{w}))$, there exists k such that $(t_k(\bar{w}), \bar{w}) \in B_{tw}(\bar{w})$ and $(\tilde{x}(t_k(\bar{w}), \bar{w}), \tilde{y}(t_k(\bar{w}), \bar{w})) \in B_{xy}(\bar{w})$. Then by the continuity of $(\tilde{x}(t, w), \tilde{y}(t, w))$, there exists a small open neighborhood $\mathcal{U}$ such that $(t_k(\bar{w}), \bar{w}) \in \mathcal{U} \subset B_{tw}(\bar{w})$ and $(\tilde{x}, \tilde{y})(\mathcal{U}) \subset B_{xy}(\bar{w})$. Since $(\tilde{x}(t, w), \tilde{y}(t, w), t, w)$ is also a solution of the system $\Phi_{\mathcal{L}}(\tilde{x}, \tilde{y}, t, w) = 0$, then by the uniqueness of $(x_{\bar{w}}(t, w), y_{\bar{w}}(t, w))$, we have $(x_{\bar{w}}(t, w), y_{\bar{w}}(t, w)) = (\tilde{x}(t, w), \tilde{y}(t, w))$ for $(t, w) \in \mathcal{U}$. Moreover, by the analyticity of $(x_{\bar{w}}(t, w), y_{\bar{w}}(t, w))$ and $(\tilde{x}(t, w), \tilde{y}(t, w))$, this equality extends to $B_{tw}(\bar{w}) \cap \{(t > 0, w > 0)\}$.

For any two different $w_1 > 0$ and $w_2 > 0$, if $B_{tw}(w_1) \cap B_{tw}(w_2) \neq \emptyset$, then $B_{tw}(w_1) \cap B_{tw}(w_2) \cap \{t > 0, w > 0\} \neq \emptyset$. Since $(x_{w_1}(t, w), y_{w_1}(t, w)) = (\tilde{x}(t, w), \tilde{y}(t, w)) = (x_{w_2}(t, w), y_{w_2}(t, w))$ on the open set $B_{tw}(w_1) \cap B_{tw}(w_2) \cap \{t > 0, w > 0\}$, by their analyticity, we must have $(x_{w_1}(t, w), y_{w_1}(t, w)) = (x_{w_2}(t, w), y_{w_2}(t, w))$ on $B_{tw}(w_1) \cap B_{tw}(w_2)$.

Therefore we are able to analytically extend $(\tilde{x}(t, w), \tilde{y}(t, w))$ to the open set $\mathcal{P} = \{t > 0, w > 0\} \bigcup (\cup_{w>0} B_{tw}(w))$, which obviously contains $\{t \geq 0, w > 0\}$.

Since $(x(t, w), y(t, w)) = (\tilde{x}_B(t, w),\ t\tilde{x}_J(t, w),\ t^2\tilde{x}_N(t, w),\ t^2\tilde{y}_B(t, w),\ t\tilde{y}_J(t, w), \tilde{y}_N(t, w))$, we can do the same thing to $(x(t, w), y(t, w))$. ∎

As an immediate consequence of Theorem 5.3, we know that $\frac{\partial x}{\partial t}(t, w)$ is bounded on the compact set $\{0 \leq t \leq t_0, \|w - e\|_\infty \leq \beta\}$. Since $x^k = x(t_k, w^k)$, $x(0, w^k) = x^*$, and $\|w^k - e\|_\infty \leq \beta$, by the mean value theorem we have $\|x^k - x^*\| = O(t_k)$.

Finally we are ready to prove the fast local convergence of Algorithm 1 in terms of t_k.

THEOREM 5.4. *Under Assumption* 2, *when* k *is sufficiently large, we have* $t_{k+1} = t_k^\theta$.

Proof. Recall that in step 1 Algorithm 1, we pick an integer $L \geq 3$ and a real number $\theta \in (1, \frac{L}{2})$.

From Theorem 5.3 and the definitions of $u(\gamma,t,w)$ and $(\hat{x}(\gamma,t,w),\hat{y}(\gamma,t,w))$, it is clear that we can analytically extend $(\hat{x}(\gamma,t,w),\hat{y}(\gamma,t,w))$ to the open set

$$\mathcal{O}=\{(\gamma,t,w)\,|(\gamma,u(\gamma,t,w))\in\mathcal{P},t>-1\},$$

which contains

$$\left\{(\gamma,t,w)\,\middle|\,w>0,t\geq 0,0\leq\gamma<t+\frac{(1+t)\min\{w_i\}}{1+\min\{w_i\}}\right\}.$$

Therefore $R(s,t,w)$ is an analytic vector function on $\{(s,t,w)|s\in R,t\geq 0,w>0\}$ and thus bounded on the compact set $\{0\leq s\leq t,0\leq t\leq t_0,\|w-e\|_\infty\leq\beta\}$. In other words, there exists a constant $\chi>0$ such that $\|R(s,t,w)\|_\infty\leq\chi$ when (s,t,w) lies in this compact set.

Since $1<\theta<\frac{L}{2}$, we have

$$t_k\leq\min\left\{2^{\frac{1}{1-\theta}},\left(\frac{\beta}{2\chi}\right)^{\frac{1}{L-2\theta}},1\right\}\Longrightarrow\sqrt{\frac{2\chi {t_k}^L}{\beta}}\leq {t_k}^\theta\leq\frac{t_k}{2}.$$

Therefore when t_k satisfies the above inequality, for all $s\in[t_k^\theta,t_k]$, by Lemma 3.3 we have

$$\begin{aligned}\|F^k(s)g^k(s)-s^2e\|_\infty-\beta s^2&\leq-(t_k-s)\left(\frac{\beta s^2}{1+t_k}-(t_k-s)^L\|R(s,t_k,w_k)\|_\infty\right)\\&\leq-(t_k-s)\left(\frac{\beta t_k^{2\theta}}{2}-\chi t_k^L\right)\\&=-\frac{\beta(t_k-s)}{2}\left(t_k^{2\theta}-\frac{2\chi t_k^L}{\beta}\right)\\&\leq 0.\end{aligned}$$

Then using the continuity of $(f^k(s),g^k(s))$ and the fact that $(f^k(t_k),g^k(t_k))=(x^k,y^k)>0$ we can show that $(f^k(s),g^k(s))>0$ for all $s\in[t_k^\theta,t_k]$.

So $t_{k+1}=t_k^\theta$. □

Remark. L denotes the number of linear systems (with the same coefficient matrix) we would like to solve in each predictor step. The bigger L is, the bigger γ can be chosen and the faster the local convergence will be. For example, when $L=3$, we can choose $\gamma=1.4$; while when $L=5$, we can choose $\gamma=2.4$.

6. Concluding remarks. In this paper we present a high-order path following method for locating the least 2-norm solution of monotone LCPs. The algorithm was partly motivated by the method proposed in [23] on finding the least 2-norm solution of linear programs. We proved the global convergence of our algorithm under the assumption that the LCP has at least one solution. We then showed the superlinear rate of convergence under the further assumption that the least 2-norm solution is maximally complementary.

Two directions are worthy of future research. The strength of this algorithm is on its superior local rate of convergence. But the high-order approximation is essentially a local technique. When the current estimate is far from the least 2-norm solution, it is usually better not to spend too much effort trying to approximate the path very

well. In order to make an efficient general algorithm, some fast globally convergent techniques have to be combined with the high-order idea.

On the other hand, the local convergence analysis itself also needs further improvement. Under the maximal complementarity assumption, we showed that $x(t,w)$ is an analytic function on $\{(t,w)|t \geq 0, w > 0\}$, and thus $\|x^k - x^*\| = \|x(t_k, w^k) - x(0, w^k)\| = O(t)$. Therefore when t_k is sufficiently small, it is generally safe to say that x^k is close to x^*. However, if the assumption does not hold, then we do not know whether $x(t,w)$ can still be analytically extended to $t = 0$. Hence it could happen that even though t is very small, x^k is still a little far from x^*. Thus it is natural and important to consider the possibility of removing the maximal complementarity assumption. Define $O_B := \{i \in B | x_i^* = 0\}$ and $O_N := \{i \in N | y_i^* = 0\}$. Then x^* satisfying the maximal complementarity condition is equivalent to $O_B = O_N = \emptyset$. Now assume $O_B \neq \emptyset$ and/or $O_N \neq \emptyset$. Based on our analysis, it is important to find an accurate bound for $x_i(t,w)$ for $i \in O_B, O_N$ as $t \longrightarrow 0+$. Some numerical examples seem to suggest that $x_i(t,w) = \Theta(t^{\frac{1}{2}})$ for $i \in O_B$, while $x_i(t,w) = \Theta(t^{\frac{3}{2}})$ for $i \in O_N$. If such bounds can be found, then we may be able to carry out the local analysis without assuming x^* to be maximally complementary, probably with a different parametrization.

Acknowledgment. The author thanks two anonymous referees for their helpful comments and suggestions.

REFERENCES

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

[2] M. Halická, *Two simple proofs for analyticity of the central path in linear programming*, Oper. Res. Lett., 28 (2001), pp. 9–19.

[3] M. Halická, *Analyticity of the central path at the boundary point in semidefinite programming*, European J. Oper. Res., 143 (2002), pp. 311–324.

[4] Z. H. Huang, L. Qi, and D. Sun, *Sub-quadratic convergence of a smoothing Newton algorithm for the P_0- and monotone LCP*, Math. Program., 99 (2004), pp. 423–441.

[5] C. Kanzow, H. Qi, and L. Qi, *On the minimum norm solution of linear programs*, J. Optim. Theory Appl., 116 (2003), pp. 333–345.

[6] M. Kojima, N. Megiddo, and T. Noma, *Homotopy continuation methods for nonlinear complementarity problems*, Math. Oper. Res., 16 (1991), pp. 754–774.

[7] O. L. Mangasarian, *Least norm solution of nonmonotone linear complementarity problems*, in Functional Analysis, Optimization, and Mathematical Economics, Oxford University Press, New York, 1990, pp. 217–221.

[8] O. L. Mangasarian and J. Ren, *New improved error bounds for the linear complementarity problem*, Math. Programming, 66 (1994), pp. 241–255.

[9] O. L. Mangasarian, *A Newton method for linear programming*, J. Optim. Theory Appl., 121 (2004), pp. 1–18.

[10] K. McShane, *Superlinearly convergent $O(\sqrt{n}L)$-iteration interior-point algorithms for linear programming and the monotone linear complementarity problem*, SIAM J. Optim., 4 (1994), pp. 247–261.

[11] S. Mizuno, *A superlinearly convergent infeasible-interior-point algorithm for geometrical LCPs without a strictly complementarity condition*, Math. Oper. Res., 21 (1996) pp. 382–400.

[12] F. A. Potra and R. Sheng, *Superlinearly convergent infeasible-interior-point algorithm for degenerate LCP*, J. Optim. Theory Appl., 97 (1998), pp. 249–269.

[13] M. Preiss and J. Stoer, *Analysis of infeasible-interior-point paths arising with semidefinite linear complementarity problems*, Math. Program., 99 (2004), pp. 499–520.

[14] J. Stoer, *High order long-step methods for solving linear complementarity problems*, Ann. Oper. Res., 103 (2001), pp. 149–159.

[15] J. Stoer and M. Wechs, *On the analyticity properties of infeasible-interior-point paths for monotone linear complementarity problems*, Numer. Math., 81 (1999), pp. 631–645.

[16] J. Stoer, M. Wechs, and S. Mizuno, *High order infeasible-interior-point methods for solving sufficient linear complementarity problems*, Math. Oper. Res., 23 (1998), pp. 832–862.

[17] J. F. Sturm, *Superlinear convergence of an algorithm for monotone linear complementarity problems, when no strictly complementary solution exists*, Math. Oper. Res., 24 (1999), pp. 72–94.

[18] P. K. Subramanian, *A note on least two norm solutions of monotone complementarity problems*, Appl. Math. Lett., 1 (1988), pp. 395–397.

[19] S. J. Wright and Y. Zhang, *A superquadratic infeasible-interior-point method for linear complementarity problems*, Math. Programming, 73 (1996), pp. 269–289.

[20] Y. Ye and K. Anstreicher, *On quadratic and $O(\sqrt{n}L)$ convergence of a predictor-corrector algorithm for LCP*, Math. Programming, 62 (1993), pp. 537–551.

[21] Y. B. Zhao and D. Li, *On a new homotopy continuation trajectory for nonlinear complementarity problems*, Math. Oper. Res., 26 (2001), pp. 119–146.

[22] Y.-B. Zhao and D. Li, *Existence and limiting behavior of a non–interior-point trajectory for nonlinear complementarity problems without strict feasible condition*, SIAM J. Control Optim., 40 (2001), pp. 898–924.

[23] Y.-B. Zhao and D. Li, *Locating the least 2-norm solution of linear programs via a path-following method*, SIAM J. Optim., 12 (2002), pp. 893–912.

[24] Y.-B. Zhao and D. Li, *A globally and locally superlinearly convergent non–interior-point algorithm for P_0 LCPs*, SIAM J. Optim., 13 (2003), pp. 1195–1221.

[25] G. Y. Zhao and J. Sun, *On the rate of local convergence of high-order-infeasible-path-following algorithms for P_*-linear complementarity problems*, Comput. Optim. Appl., 14 (1999), pp. 293–307.

SIAM J. OPTIM.
Vol. 18, No. 4, pp. 1436–1466

EXPLICIT REFORMULATIONS FOR ROBUST OPTIMIZATION PROBLEMS WITH GENERAL UNCERTAINTY SETS*

IGOR AVERBAKH† AND YUN-BIN ZHAO‡

Abstract. We consider a rather general class of mathematical programming problems with data uncertainty, where the uncertainty set is represented by a system of convex inequalities. We prove that the robust counterparts of this class of problems can be reformulated equivalently as finite and explicit optimization problems. Moreover, we develop simplified reformulations for problems with uncertainty sets defined by convex homogeneous functions. Our results provide a unified treatment of many situations that have been investigated in the literature and are applicable to a wider range of problems and more complicated uncertainty sets than those considered before. The analysis in this paper makes it possible to use existing continuous optimization algorithms to solve more complicated robust optimization problems. The analysis also shows how the structure of the resulting reformulation of the robust counterpart depends both on the structure of the original nominal optimization problem and on the structure of the uncertainty set.

Key words. robust optimization, data uncertainty, mathematical programming, homogeneous functions, convex analysis

AMS subject classifications. 90C30, 90C15, 90C34, 90C25, 90C05

DOI. 10.1137/060650003

1. Introduction.

In classical optimization models, the data are usually assumed to be known precisely. However, there are numerous situations where the data are inexact/uncertain. In many applications, the optimal solution of the nominal optimization problem may not be useful because it may be highly sensitive to small changes of the parameters of the problem.

Sensitivity analysis and stochastic programming are two traditional methods to deal with uncertain optimization problems. The former offers only local information near the nominal values of the data, while the latter requires one to make assumptions about the probability distribution of the uncertain data which may not be appropriate. Moreover, the stochastic programming approach often leads to very large optimization problems and cannot guarantee satisfaction of certain hard constraints, which is required in some practical settings.

An increasingly popular approach to optimization problems with data uncertainty is *robust optimization*, where it is assumed that possible values of data belong to some well-defined *uncertainty set*. In robust optimization, the goal is to find a solution that satisfies all constraints for any possible scenario from the uncertainty set and optimizes the worst-case (guaranteed) value of the objective function. See, e.g., [5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 21, 22, 23, 24, 25, 26, 29, 35, 39, 40]. The solutions of robust optimization models are "uniformly good" for realizations of data from the

*Received by the editors January 15, 2006; accepted for publication (in revised form) June 26, 2007; published electronically January 16, 2008.
http://www.siam.org/journals/siopt/18-4/65000.html

†Division of Management, University of Toronto at Scarborough, Scarborough, ON, M1C 1A4, Canada (averbakh@utsc.utoronto.ca). The research of this author was supported by a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC).

‡Institute of Applied Mathematics, AMSS, Chinese Academy of Sciences, Beijing 100080, China (ybzhao@amss.ac.cn), and Division of Management, University of Toronto at Scarborough, ON, M1C 1A4, Canada. The research of this author was supported by grants 10671199 and 70221001 from the National Natural Science Foundation of China and partially supported by CONACyT-SEP project SEP-2004-C01-45786, Mexico.

uncertainty set. Early work in this direction was done by Soyster [39, 40] and Falk [22] under the name of "inexact linear programming." The robust optimization approach has been applied to various problems in operations management, financial planning, and engineering design (see, e.g., [29, 26, 10, 6, 31, 35]).

A formulation of a robust model as a mathematical programming problem is called a *robust counterpart*. Since in the robust approach the constraints must be satisfied for all possible realizations of data from the uncertainty set, the robust counterpart is typically a complicated semi-infinite optimization problem. A fundamental question in robust optimization is whether the robust counterpart can be represented as a single finite and explicit optimization problem, so that existing optimization methods can be used to solve it. Such an analysis also helps to understand computational complexity of robust optimization problems.

So far, to obtain sufficiently simple robust counterparts, the uncertainty set was normally assumed to have a fairly simple structure, for example, a Cartesian product of intervals, an ellipsoid, an intersection of ellipsoids, or a set defined by certain norms (see, e.g., [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 23, 24, 25, 26, 29]). Of course, the simpler the uncertainty set is, the easier it is to solve the robust optimization problem, and in some situations simplifying assumptions about uncertainty sets is natural when modelling a practical problem. However, more complicated uncertainty sets may be encountered in both theoretical study and in applications (see Remark 3.1 for details). Therefore, it is important to understand possibilities of the robust approach dealing with problems involving complicated or general uncertainty sets. Study of robust optimization problems with general uncertainty sets may provide additional tools for modelling intricate real-life situations and a unified treatment of specialized cases. Moreover, such a study can provide additional insights and results and even improve known results for some specialized cases when general results are reduced to such specialized cases (see section 6 for details).

In this paper, we consider robust optimization problems with uncertainty sets defined by a system of convex inequalities. The optimization problems we consider may be nonconvex and are wide enough to include linear programming, linear complementarity problems, quadratic programming, second order cone programming, and general polynomial programming problems. We prove that the robust counterparts of the considered problems with uncertainty are finite optimization problems which can be formulated by using the nominal data of the underlying optimization problem and the conjugates of the functions defining the uncertainty set. Compared with the original optimization problem, a major extra difficulty of the robust counterpart comes from the conjugates of the functions that define the uncertainty set. The conjugates of these functions usually are not given explicitly and may be difficult to compute. To identify explicit and simplified formulations of robust counterparts, we focus on a class of convex functions whose conjugates can be expressed explicitly. Our strongest results and simplest reformulations of robust counterparts correspond to the case where the uncertainty sets are defined by convex homogeneous functions. This class of uncertainty sets is broad enough to include most uncertainty models that have been investigated in the literature, as well as many other important cases, for example, where deviations of data from nominal values may be asymmetric and not even defined by norms.

We note that instead of optimizing the worst-case value of the objective function, another possibility is to optimize the worst-case regret, which is the worst-case deviation of the objective function value from the optimal value under the realized scenario,

or, in other words, to minimize the worst-case loss in the objective function value that may occur because the decision is made before the realized scenario is known. This criterion leads to minmax regret optimization models [29, 2, 1, 3, 4]. Minmax regret problems are typically computationally hard [29, 4], although there are exceptions (see, e.g., [2, 1, 3]). Minmax regret problems also fit the general paradigm of robust optimization, but we do not consider them in this paper. We also note that there are other concepts of robustness in the literature under the name of "model uncertainty" or "ambiguity." See, e.g., [42, 28, 17, 33, 18, 38, 27, 37, 16, 19, 20, 21, 31, 41].

This paper is organized as follows. In section 2, we describe the class of optimization problems that we consider. In section 3, we define the uncertainty set of data and provide an equivalent, deterministic representation of the robust optimization problems via Fenchel's conjugate functions. In section 4, we give an explicit representation for the robust counterpart when the uncertainty set is defined by (nonhomogeneous) convex functions that fall in the linear space generated by homogeneous functions of arbitrary degrees. The case of uncertainty sets defined by homogeneous functions is studied in section 5. Specializing the general results of sections 3, 4, and 5 to robust problems where the nominal problem is a linear programming problem and/or the uncertainty set is of a special type commonly used in the literature is discussed in section 6, and concluding remarks are provided in section 7.

2. A class of optimization problems with data uncertainty. We consider the following optimization problem:

$$\min\{c^T x : \ f_i(x) \le b_i, \ i = 1, \dots, m, \ F(x) \le 0\}, \tag{1}$$

where $c = (c_1, \dots, c_n)^T$ and $b = (b_1, \dots, b_m)^T$ are fixed vectors, and f_i's are functions of the form

$$f_i(x) = \left(W^{(i)}(x)\right)^T M^{(i)} V^{(i)}(x), \quad i = 1, \dots, m, \tag{2}$$

where $W^{(i)}(x)$ and $V^{(i)}(x)$ are two mappings from R^n to R^{N_i}, $M^{(i)}$ is an $N_i \times N_i$ real matrix, and N_i's are positive integers. We write $W^{(i)}(x)$ and $V^{(i)}(x)$ as $W^{(i)}(x) = (W_1^{(i)}(x), \dots, W_{N_i}^{(i)}(x))^T$ and $V^{(i)}(x) = (V_1^{(i)}(x), \dots, V_{N_i}^{(i)}(x))^T$, where each $W_j^{(i)}$ $(j = 1, \dots, N_i)$ is a function from R^n to R.

We assume that only the data $M^{(i)}$, $i = 1, \dots, m$, are subject to uncertainty. In (1), $F(x) \le 0$ denotes constraints without uncertainty, e.g., the simple constraints $x \ge 0$. We assume that c and b are certain without loss of generality, because a problem with uncertain c and b can be easily transformed into a problem with certain coefficients of the objective function and right-hand sides of the constraints. Also, if the objective function is not linear, it can be made linear by introducing an additional variable and a new constraint. We note that functions f_i are linear in the uncertain data $M^{(i)}$ (but can be nonlinear in the decision variables x).

The above optimization model is very general. For example, it includes the following important special cases.

Linear programming (LP). Let $A \in R^{m \times n}$ (i.e., an $m \times n$ matrix) and $b = (b_1, \dots, b_m)^T$. Without loss of generality, we assume $m \le n$. Consider functions $f_i(x)$ of the form (2), where

$$W^{(i)}(x) = e_i \in R^n, \quad V^{(i)}(x) = x \in R^n, \quad M^{(i)} = \begin{bmatrix} A \\ 0 \end{bmatrix}_{n \times n},$$

where e_i, throughout this paper, denotes the ith column of an $n \times n$ identity matrix, and 0 in $M^{(i)}$ denotes an $(n-m) \times n$ zero matrix. It is evident that the inequalities $f_i = (W^{(i)})^T M^{(i)} V^{(i)} \leq b_i,\ i = 1, \ldots, m$, are equivalent to $Ax \leq b$. Therefore, problem (1) with $F(x) = -x \leq 0$ reduces to the following LP problem:

$$\min\{c^T x : \ Ax \leq b,\ x \geq 0\}. \tag{3}$$

This implies that the LP problem (3) with uncertain coefficient matrix A is a special case of the optimization problem (1) with uncertain data $M^{(i)}$. There is also another way to write an LP problem in the form (1)–(2); see (38) and (39) in section 6.2 for details.

Linear complementarity problem (LCP). Given a matrix $M \in R^{n \times n}$ and a vector $q \in R^n$, the LCP is defined as

$$Mx + q \geq 0, \quad x \geq 0, \quad x^T(Mx + q) = 0.$$

Solutions to the LCP are very sensitive to changes in data because of the equation $x^T(Mx + q) = 0$. When the matrix M is uncertain, it is hard to find a solution that satisfies the above system and is "immune" to changes of M. Thus, it is reasonable to consider the optimization form of the LCP, i.e.,

$$\min\{x^T(Mx + q) : \ Mx + q \geq 0,\ x \geq 0\},$$

or equivalently

$$\min\{t : \ x^T(Mx + q) - t \leq 0,\ Mx + q \geq 0,\ x \geq 0\},$$

which is less sensitive in the sense that it is equivalent to the LCP if the LCP has a solution and can still have a solution even when the LCP has no solution. The above optimization problem can be reformulated as (2) by letting

$$W^{(1)}(x) = \begin{pmatrix} x \\ 1 \\ e_1 \end{pmatrix} \in R^{2n+1}, \quad W^{(i)} = \begin{pmatrix} 0^{(n+1)} \\ e_{i-1} \end{pmatrix} \in R^{2n+1} \quad \text{for } i = 2, \ldots, n+1,$$

$$M^{(i)} = \begin{bmatrix} M & q & 0 & \overbrace{0 \ \ldots 0}^{n-1} \\ 0 & 0 & -1 & 0 \ \ldots 0 \\ -M & -q & 0 & 0 \ \ldots 0 \end{bmatrix},$$

$$V^{(i)}(x) = \begin{pmatrix} x \\ 1 \\ t \\ 0^{(n-1)} \end{pmatrix} \in R^{2n+1}, \quad i = 1, \ldots, n+1,$$

where $t \in R$, and $0^{(n+1)}$ and $0^{(n-1)}$ denote $(n+1)$- and $(n-1)$-dimensional zero vectors, respectively. It is easy to verify that problem (1) with $F(x) = -x \leq 0$ and $f_i = (W^{(i)})^T M^{(i)} V^{(i)} \leq 0$ $(i = 1, \ldots, n+1)$ is the same as the optimization form of the LCP. It is worth mentioning that Zhang [43] considered equality constrained robust optimization, and his approach may be also used to deal with LCPs with uncertain data.

(Nonconvex) quadratic programming (QP). Consider functions $f_i(x)$ of the form (2), where

$$W^{(i)}(x) = \begin{pmatrix} x \\ 1 \end{pmatrix} \in R^{n+1} \quad \text{for } i = 0, \ldots, m,$$

$$V^{(0)}(x) = \begin{pmatrix} x \\ t \end{pmatrix} \in R^{n+1}, \; V^{(i)}(x) = \begin{pmatrix} x \\ 0 \end{pmatrix} \in R^{n+1} \quad \text{for } i = 1, \ldots, m$$

and

$$M^{(i)} = \begin{bmatrix} Q_i & 0 \\ q_i^T & -1 \end{bmatrix}_{(n+1)\times(n+1)} \qquad \text{for } i = 0, \ldots, m, \tag{4}$$

where each Q_i is an $n \times n$ symmetric matrix and each q_i is a vector in R^n. Then the optimization problem (1) with the objective t and constraints $f_i = (W^{(i)})^T M^{(i)} V^{(i)} \leq -c_i (i = 0, \ldots, m)$ is reduced to the following QP problem:

$$\begin{aligned} \min \quad & x^T Q_0 x + q_0^T x + c_0 \\ \text{s.t.} \quad & x^T Q_i x + q_i^T x + c_i \leq 0 \quad \text{for } i = 1, \ldots, m. \end{aligned}$$

Thus, a QP problem with uncertain coefficients $(Q_i, q_i)(i = 0, \ldots, m)$ can be represented as an optimization problem (1) with uncertain data $M^{(i)}$ given as (4).

Second order cone programming (SOCP). Let $A \in R^{m\times n}, b \in R^m, c \in R^n$, and β be a scalar. Let

$$W^{(1)}(x) = V^{(1)}(x) = \begin{pmatrix} x \\ 1 \end{pmatrix} \in R^{n+1}$$

and

$$M^{(1)} = \begin{bmatrix} A^T A - cc^T & 0 \\ 2b^T A - 2\beta c^T & b^T b - \beta^2 \end{bmatrix}, \tag{5}$$

and $W^{(2)}(x) = e \in R^n$ (the vector with all components equal to 1), $V^{(2)}(x) = x \in R^n$, and

$$M^{(2)} = \begin{bmatrix} -c^T \\ 0 \end{bmatrix}_{n\times n}.$$

Then the constraint $f_1 = (W^{(1)})^T M^{(1)} V^{(1)} \leq 0$, together with $f_2 = (W^{(2)})^T M^{(2)} V^{(2)} \leq \beta$, is equivalent to the second order cone constraint: $\|Ax + b\| \leq c^T x + \beta$. In fact, $f_1 \leq 0$ can be written as

$$(Ax + b)^T (Ax + b) \leq (c^T x + \beta)^2,$$

and $f_2 \leq \beta$ can be written as $c^T x + \beta \geq 0$. Combination of these two inequalities leads to a second order cone constraint. Thus, uncertainty of the data (A, B, c, β) leads to uncertainty of the matrices $M^{(1)}$ and $M^{(2)}$.

Polynomial programming. We recall that a monomial in $x_1, \ldots, x_n$ is a product of the form $x_1^{\alpha_1} \cdot x_2^{\alpha_2} \cdots x_n^{\alpha_n}$, where $\alpha_1, \ldots, \alpha_n$ are nonnegative integers. It is evident

that if the components of $W(x)$ and $V(x)$ are monomials, then for any given matrix M, a function of the form (2) is a polynomial. Conversely, any real polynomial is a linear combination of some monomials, i.e.,

$$P(x_1, x_2, \ldots, x_n) = \sum_{(\alpha_1,\alpha_2,\ldots,\alpha_n)} C^{(\alpha_1,\alpha_2,\ldots,\alpha_n)} x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_n^{\alpha^n},$$

where $C^{(\alpha_1,\ldots,\alpha_n)}$ are real coefficients. Then the simplest way to write it in the form (2) is to set $W(x) = e$, set $V(x)$ to be the vector of all monomials $x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_n^{\alpha^n}$ appearing in $P(x)$, and set M to be the diagonal matrix with diagonal entries $C^{(\alpha_1,\alpha_2,\ldots,\alpha_n)}$. Thus polynomial optimization with uncertain coefficients is a special case of (1) with uncertain data $M^{(i)}$.

3. Robust counterparts as finite deterministic optimization problems. We start with a description of the uncertainty set. *Let K_i, $i = 1, \ldots, m$, be a bounded subset of $R^{N_i^2}$ that contains the origin.* Suppose that the uncertain data $M^{(i)}$ $(i = 1, \ldots, m)$ of the ith constraint of (1) are allowed to vary in such a way that the deviations from their fixed nominal values $\overline{M}^{(i)}$ fall in K_i. That is, the uncertainty set of the data $M^{(i)}$ is defined as

$$\mathcal{U}_i = \left\{ \widetilde{M}^{(i)} \,\middle|\, vec(\widetilde{M}^{(i)}) - vec(\overline{M}^{(i)}) \in K_i \right\}, \quad i = 1, \ldots, m, \tag{6}$$

where for a given matrix M, $vec(M)$ denotes the vector obtained by stacking the transposed rows of M on top of one another. Then the robust counterpart of the optimization problem (1) with uncertainty sets $\mathcal{U}_i$ is defined as follows:

$$\min c^T x$$

$$\text{s.t. } f_i = \left(W^{(i)}(x)\right)^T \widetilde{M}^{(i)} V^{(i)}(x) \le b_i \quad \forall \widetilde{M}^{(i)} \in \mathcal{U}_i, i = 1, \ldots, m, F(x) \le 0, \tag{7}$$

which is a semi-infinite optimization problem. The optimal solution to this problem is feasible for all realizations of the data $\widetilde{M}^{(i)}$.

We denote by $\delta(u|K)$ the indicator function of a set K (see [36]), and the conjugate function of $\delta(u|K)$ is denoted by $\delta^*(u|K)$, which is equal to the support function $\psi_K(u) = \max\{u^T v : v \in K\}$. First we state the following general result, which shows that the robust counterpart (7) can be written equivalently as a finite deterministic optimization problem, regardless of the type of uncertainty sets.

THEOREM 3.1. *The robust optimization problem* (7) *is equivalent to the following finite and deterministic optimization problem:*

$$\begin{aligned} &\min c^T x \\ &s.t. \left(W^{(i)}(x)\right)^T \overline{M}^{(i)} V^{(i)}(x) + \delta^*(\chi_i | \mathrm{cl}(\mathrm{co} K_i)) \le b_i, \quad i = 1, \ldots, m, \\ &\quad F(x) \le 0, \end{aligned}$$

where $\mathrm{cl}(\mathrm{co}K_i)$ *denotes the closure of the convex hull of set* K_i, *and* $\chi_i = W^{(i)}(x) \otimes V^{(i)}(x) \in R^{N_i^2}$, *i.e., is the Kronecker product of the vectors* $W^{(i)}(x)$ *and* $V^{(i)}(x)$.

Proof. In fact, the constraint $f_i = (W^{(i)}(x))^T \widetilde{M}^{(i)} V^{(i)}(x) \le b_i$ for all $vec(\widetilde{M}^{(i)}) - vec(\overline{M}^{(i)}) \in K_i$ is equivalent to

$$\sup \left\{ W^{(i)}(x)^T \widetilde{M}^{(i)} V^{(i)}(x) : vec(\widetilde{M}^{(i)}) - vec(\overline{M}^{(i)}) \in K_i \right\} \le b_i. \tag{8}$$

Notice that for any square matrices B, C, we have $\mathrm{tr}(BC) = (vec(B))^T vec(C^T)$. Thus, we have

$$\begin{aligned}\left(W^{(i)}(x)\right)^T \widetilde{M}^{(i)} V^{(i)}(x) &= \mathrm{tr}\left(\widetilde{M}^{(i)} V^{(i)}(x) \left(W^{(i)}(x)\right)^T\right) \\ &= \left(vec(\widetilde{M}^{(i)})\right)^T vec\left(W^{(i)}(x)\left(V^{(i)}(x)\right)^T\right) \\ &= \left(vec(\widetilde{M}^{(i)})\right)^T \left(W^{(i)}(x) \otimes V^{(i)}(x)\right).\end{aligned}$$

Denoting $\chi_i = W^{(i)}(x) \otimes V^{(i)}(x)$, the constraint (8) can be written as

$$\begin{aligned} b_i &\geq \sup\left\{\left(vec(\widetilde{M}^{(i)})\right)^T \chi_i : vec(\widetilde{M}^{(i)}) - vec(\overline{M}^{(i)}) \in K_i\right\} \\ &= \left(vec(\overline{M}^{(i)})\right)^T \chi_i + \sup_{u \in K_i} u^T \chi_i = \left(vec(\overline{M}^{(i)})\right)^T \chi_i + \sup_{u \in \mathrm{cl}(\mathrm{co}K_i)} u^T \chi_i \\ &= \left(W^{(i)}(x)\right)^T \overline{M}^{(i)} V^{(i)}(x) + \delta^*(\chi_i | \mathrm{cl}(\mathrm{co}K_i)).\end{aligned}$$

The original semi-infinite constraints become finite and deterministic constraints. □

For robust optimization, when the uncertainty set is not convex, the robust counterpart remains unchanged if we replace the uncertainty set by its closed convex hull. This observation was first mentioned in [7] and can be seen clearly from the above result. Because of this fact, we may assume without loss of generality that each K_i is a closed convex set. In applications, the convex set K_i is usually determined by a system of convex inequalities. So, *throughout the rest of the paper, we assume that K_i is a closed, bounded convex set containing the origin and it can be represented as*

$$K_i = \left\{u \;\middle|\; g_j^{(i)}(u) \leq \Delta_j^{(i)},\ j = 1, \ldots, \ell^{(i)}\right\}, \quad i = 1, \ldots, m, \tag{9}$$

where $\ell^{(i)}$'s are given integers, $\Delta_j^{(i)}$'s are constants, and $g_j^{(i)}$'s are proper closed convex functions from $R^{N_i^2}$ to $\overline{R}$. Here $\overline{R} = R \cup \{+\infty\}$ and "proper" means that the function is finite somewhere (throughout the paper, we use the terminology from [36]). Since $0 \in K_i$, we have $g_j^{(i)}(0) \leq \Delta_j^{(i)}$ for all $j = 1, \ldots, \ell^{(i)}$.

Remark 3.1. In this remark, we give additional motivation for considering the general uncertainty set (9) as opposed to special uncertainty sets studied in the literature. We note that the importance of studying robust problems with complicated uncertainty sets was emphasized, for example, in [15].

(i) Consider the following uncertainty set:

$$\mathcal{U} = \left\{D \;\middle|\; \exists z \in R^{|N|} : D = D_0 + \psi(z) = D_0 + \sum_{j \in N} \Delta D_j z_j,\ \|z\| \leq \Omega\right\}, \tag{10}$$

where Ω is a given number, D_0 is a given vector (nominal values of the uncertain data), and ΔD_j's are directions of data perturbation. This uncertainty set has been widely used in the literature (see, e.g., [5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 23, 24, 25, 26].

It is the image of a ball (defined by some norm) under linear transformation; i.e., the function $\psi(z)$ here is a linear function in z. This widely used uncertainty set can be written in the form (9) with only one convex inequality $g(u) \leq \Omega$, where function $g(u)$ is also homogeneous of 1-degree, and $g(u)$ is not a norm, in general, unless $|N|$ is equal to the number of data and the data perturbation directions ΔD^j's are linearly independent (see section 6.1 for details). This typical example shows that it is necessary to study the case when the functions $g_j^{(i)}(u)$ in (9) are convex and homogeneous (but not necessarily norms). Section 5 of this paper is devoted to this important case.

For the uncertainty set $\mathcal{U}$ defined by (10), the function $\psi(z)$ is linear in z. In some applications, however, such a model is insufficient for description of more complicated uncertainty sets. The next two examples show that in some situations the function $\psi(z)$ may be nonlinear, and hence the uncertainty set may be much more complicated.

(ii) Consider SOCP. It is often assumed that the data (A, b, c) are subject to an ellipsoidal uncertainty set which is the case of (10) where the norm is the 2-norm. When we reformulate SOCP into the form of (1), the data $M^{(1)}$ is determined by the matrix (5). It is easy to see the data $M^{(1)}$ belongs to the following uncertainty set:

$$\mathcal{U} = \left\{ D \,\middle|\, \exists z \in R^{|N|} : D = D^0 + \psi(z), \quad \|z\| \leq \Omega \right\}, \tag{11}$$

where $\psi(z)$ is a quadratic function in z. Thus, this example shows that a more complicated uncertainty set than (10) might appear when we make a reformulation of the problem. Such reformulations are often made when a problem is studied from different perspectives.

(iii) This example, taken from [23], shows that a nonlinear function $\psi(z)$ arises in (11) when robust interpolation problems are considered. Let $n \geq 1$ and k be given integers. We want to find a polynomial of degree $n-1$, $p(t) = x_1 + \cdots + x_n t^{n-1}$ that interpolates given points (a_i, b_i), i.e., $p(a_i) = b_i, i = 1, \ldots, k$. If interpolation points (a_i, b_i) are known precisely, we obtain the following linear equation:

$$\begin{bmatrix} 1 & a_1 & \cdots & a_1^{n-1} \\ \vdots & \vdots & & \vdots \\ 1 & a_k & \cdots & a_k^{n-1} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}.$$

Now assume that a_i's are not known precisely, i.e., $a_i(\delta) = a_i + \delta_i, i = 1, \ldots, k$, where the $\delta = (\delta_1, \ldots, \delta_k)$ is unknown but bounded, i.e., $\|\delta\|_\infty \leq \rho$, where $\rho \geq 0$ is given. A robust interpolant is a solution x that minimizes $\|A(\delta)x - b\|$ over the region $\|\delta\|_\infty \leq \rho$, where

$$A(\delta) = \begin{bmatrix} 1 & a_1(\delta) & \cdots & a_1(\delta)^{n-1} \\ \vdots & \vdots & & \vdots \\ 1 & a_k(\delta) & \cdots & a_k(\delta)^{n-1} \end{bmatrix}$$

is an uncertain Vandermonde matrix. Such a matrix can be written in the form (11) with nonlinear function $\psi(z)$. In fact, we have (see [23] for details)

$$A(\delta) = A(0) + L\Delta(I - D\Delta)^{-1} R_A,$$

where L, D, and R_A are constant matrices determined by a_i's, and $\Delta = \oplus_{i=1}^k \delta_i I_{n-1}$.

(iv) Our model provides a unified treatment of many uncertainty sets in the literature. Note that (11) can be written in the form (9) by letting $g(D) = \inf\{\|z\| : D = \psi(z)\}$. Then $\mathcal{U} - \{D_0\} = \{D : \ g(D) \leq \Omega\}$. This can be proved by the same argument as Lemma 6.1.

(v) Studying problems with general uncertainty sets may in fact lead to new or stronger results for important special cases, as we demonstrate in section 6.

Since robust optimization problems, in general, are semi-infinite optimization problems which are hard to solve, the fundamental question is whether a robust optimization problem can be explicitly represented as an equivalent finite optimization problem, so that the existing optimization methods can be applied. We are addressing this question in this paper. It should be mentioned that, generally, two research directions are possible: (1) developing computationally tractable approximate (relaxed) formulations; (2) developing exact formulations which, naturally, will be computationally difficult for sufficiently complicated nominal problems and/or uncertainty sets. Our paper focuses on the second direction; the first direction was investigated, for instance, in Bertsimas and Sim [14]. We believe that both directions are important for theoretical and practical progress in robust optimization; we comment on this in more detail in section 6.

Let us mention some auxiliary results and definitions. Given a function f we denote its domain by $\mathrm{dom}(f)$ and denote its Fenchel's conjugate function by f^*, i.e.,

$$f^*(w) = \sup_{x \in \mathrm{dom}(f)} \left(w^T x - f(x)\right).$$

We recall that the infimal convolution function of $g_j (j = 1, \ldots, \ell)$, denoted by $g_1 \diamond g_2 \diamond \cdots \diamond g_\ell$, is defined as

$$(g_1 \diamond g_2 \diamond \cdots \diamond g_\ell)(u) = \inf \left\{ \sum_{j=1}^{\ell} g_j(u_j) : \sum_{j=1}^{\ell} u_j = u \right\}.$$

The following result will be used in our later analysis.

LEMMA 3.1 (see [36, Theorem 16.4]). *Let $f_1, \ldots, f_\ell : R^n \to \overline{R}$ be proper convex functions. Then $(\mathrm{cl}(f_1) + \cdots + \mathrm{cl}(f_\ell))^* = \mathrm{cl}(f_1^* \diamond \cdots \diamond f_\ell^*)$, where $\mathrm{cl}(f)$ denotes the closure of the convex function f. If the relative interiors of the domains of these functions, i.e., $ri(\mathrm{dom}(f_i)), i = 1, \ldots, \ell$, have a point in common, then*

$$\left(\sum_{i=1}^{\ell} f_i\right)^* (x) = (f_1^* \diamond \cdots \diamond f_\ell^*)(x) = \inf \left\{ \sum_{i=1}^{\ell} f_i^*(x_i) : \sum_{i=1}^{\ell} x_i = x \right\},$$

where for each $x \in R^n$ the infimum is attained.

Now we consider the robust programming problem (7) where the uncertainty set is determined by (6) and (9). We have the following general result.

THEOREM 3.2. *Let K_i $(i = 1, \ldots, m)$ be given by (9), where each $g_j^{(i)} (j = 1 \ldots, \ell^{(i)})$ is a closed proper convex function. Suppose that Slater's condition holds for each i; i.e., for each i, there exists a point $u_0^{(i)}$ such that $g_j^{(i)}(u_0^{(i)}) < \Delta_j^{(i)}$ for all $j = 1, \ldots, \ell^{(i)}$. Then the robust counterpart (7) is equivalent to*

$$\min c^T x$$

$$s.t.\ \left(W^{(i)}(x)\right)^T \overline{M}^{(i)} V^{(i)}(x) + \sum_{j=1}^{\ell^{(i)}} \lambda_j^{(i)} \Delta_j^{(i)} + \left(\sum_{j=1}^{\ell^{(i)}} \lambda_j^{(i)} g_j^{(i)}\right)^* (\chi_i) \le b_i,$$

$$i = 1, \ldots, m, \quad \lambda_j^{(i)} \ge 0, \quad j = 1, \ldots, \ell^{(i)}, \quad i = 1, \ldots, m,$$

$$F(x) \le 0,$$

where $\chi_i = W^{(i)}(x) \otimes V^{(i)}(x)$. *This problem can be further written as*

$$\min c^T x$$

$$s.t.\ \left(W^{(i)}(x)\right)^T \overline{M}^{(i)} V^{(i)}(x) + \sum_{j=1}^{\ell^{(i)}} \lambda_j^{(i)} \Delta_j^{(i)} + \Upsilon^{(i)}(\lambda^{(i)}, u^{(i)}) \le b_i, \quad i = 1, \ldots, m,$$

$$\chi_i = \begin{cases} \sum_{j \in J_i} u_j^{(i)} & \text{if } J_i \neq \emptyset, \\ 0 & \text{otherwise,} \end{cases} \quad i = 1, \ldots, m, \tag{12}$$

$$\lambda_j^{(i)} \ge 0, \quad j = 1, \ldots, \ell^{(i)}, \quad i = 1, \ldots, m,$$

$$F(x) \le 0,$$

where $J_i = \{j : \lambda_j^{(i)} > 0, j = 1, \ldots, \ell^{(i)}\}$, $\lambda^{(i)}$ *denotes the vector whose components are* $\lambda_j^{(i)}$, $j = 1, \ldots, \ell^{(i)}$, $u^{(i)}$ *denotes the vector whose components are* $u_j^{(i)}, j \in J_i$, *and*

$$\Upsilon^{(i)}(\lambda^{(i)}, u^{(i)}) = \begin{cases} \sum_{j \in J_i} \lambda_j^{(i)} \left(g_j^{(i)}\right)^* \left(u_j^{(i)} / \lambda_j^{(i)}\right) & \text{if } J_i \neq \emptyset, \\ 0 & \text{otherwise.} \end{cases}$$

Proof. We see from the proof of Theorem 3.1 that x is feasible to the robust problem (7) if and only if $F(x) \le 0$ and for each i we have

$$\left(W^{(i)}(x)\right)^T \overline{M}^{(i)} V^{(i)}(x) + \max_{u \in K_i} u^T \chi_i \le b_i. \tag{13}$$

Let $Z(\chi_i) = \max\{u^T \chi_i : u \in K_i\}$, where K_i is given by (9), which by our assumption is a bounded, closed convex set. Thus the maximum value of the convex optimization problem $\max\{u^T \chi_i : u \in K_i\}$ is finite and attainable. Denote the Lagrangian multiplier vector for this problem by $\lambda^{(i)} = (\lambda_1^{(i)}, \lambda_2^{(i)}, \ldots, \lambda_{\ell^{(i)}}^{(i)}) \in R_+^{\ell^{(i)}}$. Since Slater's condition holds for the problem $\max\{u^T \chi_i : u \in K_i\}$, by Lagrangian saddle-point theorem (see, e.g., Theorem 28.3, Corollary 28.3.1, and Theorem 28.4 in [36]), we have

$$\begin{aligned} Z(\chi_i) &= -\min\{-u^T \chi_i : \ g_j^{(i)}(u) \le \Delta_j^{(i)}, j = 1, \ldots, \ell^{(i)}\} \\ &= -\sup_{\lambda^{(i)} \in R_+^{\ell^{(i)}}} \inf_{u \in R^{N_i^2}} \left(-u^T \chi_i + \sum_{j=1}^{\ell^{(i)}} \lambda_j^{(i)} \left(g_j^{(i)}(u) - \Delta_j^{(i)}\right)\right) \\ &= -\sup_{\lambda^{(i)} \in R_+^{\ell^{(i)}}} \left[-\sum_{j=1}^{\ell^{(i)}} \lambda_j^{(i)} \Delta_j^{(i)} + \inf_{u \in R^{N_i^2}} \left(-u^T \chi_i + \sum_{j=1}^{\ell^{(i)}} \lambda_j^{(i)} g_j^{(i)}(u)\right)\right] \end{aligned} \tag{14}$$

$$= - \sup_{\lambda^{(i)} \in R_+^{\ell^{(i)}}} \left[- \sum_{j=1}^{\ell^{(i)}} \lambda_j^{(i)} \Delta_j^{(i)} - \sup_{u \in R^{N_i^2}} \left(u^T \chi_i - \sum_{j=1}^{\ell^{(i)}} \lambda_j^{(i)} g_j^{(i)}(u) \right) \right]$$

$$= - \sup_{\lambda^{(i)} \in R_+^{\ell^{(i)}}} \left(- \sum_{j=1}^{\ell^{(i)}} \lambda_j^{(i)} \Delta_j^{(i)} - \left(\sum_{i=1}^{\ell^{(i)}} \lambda_j^{(i)} g_j^{(i)} \right)^* (\chi_i) \right)$$

$$= \inf_{\lambda^{(i)} \in R_+^{\ell^{(i)}}} \left(\sum_{j=1}^{\ell^{(i)}} \lambda_j^{(i)} \Delta_j^{(i)} + \left(\sum_{j=1}^{\ell^{(i)}} \lambda_j^{(i)} g_j^{(i)} \right)^* (\chi_i) \right).$$

Under our assumptions, the above infimum is attainable (by the existence of a saddle point of the Lagrangian function [36]). Substituting (14) into (13), we see that x satisfies (13) if and only if it satisfies the following inequalities for some $\lambda^{(i)}$:

$$\left(W^{(i)}(x) \right)^T \overline{M}^{(i)} V^{(i)}(x) + \sum_{j=1}^{\ell^{(i)}} \lambda_j^{(i)} \Delta_j^{(i)} + \left(\sum_{j=1}^{\ell^{(i)}} \lambda_j^{(i)} g_j^{(i)} \right)^* (\chi_i) \le b_i, \tag{15}$$

$$\lambda^{(i)} = (\lambda_1^{(i)}, \lambda_2^{(i)}, \ldots, \lambda_{\ell^{(i)}}^{(i)}) \in R_+^{\ell^{(i)}}. \tag{16}$$

Indeed, if x is feasible to (13), since the infimum in (14) is attainable, there exists some $\lambda^{(i)} \in R_+^{\ell^{(i)}}$ such that $(x, \lambda^{(i)})$ is feasible to the system (15)–(16). Conversely, if $(x, \lambda^{(i)})$ is feasible to (15) and (16), then by (14), we see that (15) implies (13). Replacing (13) by (15) together with (16), the first part of the desired result follows from Theorem 3.1.

We now derive the optimization problem (12). Suppose that $(x, \lambda^{(i)})$ satisfies (15) and (16). We have two cases.

Case 1. $J_i = \{j : \lambda_j^{(i)} > 0, j = 1, \ldots, \ell^{(i)}\} \neq \emptyset$. Denote by $u^{(i)}$ the vector whose components are $u_j^{(i)}, j \in J_i$. Notice that for any constant $\alpha > 0$, the conjugate $(\alpha f)^*(x) = \alpha f^*(x/\alpha)$. For given $\lambda^{(i)} \in R_+^{\ell^{(i)}}$, by Lemma 3.1, we have

$$\left(\sum_{j=1}^{\ell^{(i)}} \lambda_j^{(i)} g_j^{(i)} \right)^* (\chi_i) = \inf_{u^{(i)}} \left\{ \sum_{j \in J_i} \lambda_j^{(i)} \left(g_j^{(i)} \right)^* (u_j^{(i)} / \lambda_j^{(i)}) : \chi_i = \sum_{j \in J_i} u_j^{(i)} \right\}.$$

Again, by Lemma 3.1, the infimum above is attainable, and hence there are $u_j^{(i)}$, $j \in J_i$, such that

$$\left(\sum_{j=1}^{\ell^{(i)}} \lambda_j^{(i)} g_j^{(i)} \right)^* (\chi_i) = \sum_{j \in J_i} \lambda_j^{(i)} \left(g_j^{(i)} \right)^* (u_j^{(i)} / \lambda_j^{(i)}),$$

$$\chi_i = \sum_{j \in J_i} u_j^{(i)}.$$

Case 2. $J_i = \emptyset$. Notice that

$$\left(\sum_{j=1}^{\ell^{(i)}} \lambda_j^{(i)} g_j^{(i)} \right)^* (w) = \sup_{u \in R^n} (w^T u - 0) = \begin{cases} \infty & \text{if } w \neq 0, \\ 0 & \text{if } w = 0. \end{cases}$$

Since $(x, \lambda^{(i)})$ is feasible to (15) and (16), we conclude that for this case

$$\chi_i = 0, \quad \left(\sum_{j=1}^{\ell^{(i)}} \lambda_j^{(i)} g_j^{(i)}\right)^* (\chi_i) = 0.$$

Combining the above two cases leads to the optimization problem (12). □

We see from Theorem 3.2 that the level of complexity of the robust counterpart, compared with the nominal optimization problem, is determined mainly by the conjugate functions $(g_j^{(i)})^*$ $(j = 1, \ldots, \ell^{(i)}, i = 1, \ldots, m)$ and functions χ_i $(i = 1, \ldots, m)$. The more complicated the conjugate functions are, the more difficult the robust counterpart is. Notice that the constraint $\sum_{j \in J_i} u_j^{(i)} = \chi_i$ is an explicit expression, and in some cases, e.g., LP, χ_i is linear in x and thus does not add difficulty. We also note that when $\ell^{(i)} = 1$, i.e., when K_i is defined by only one constraint, then $u_j^{(i)} = \chi_i$, in which case the formula $\sum_{j \in J_i} u_j^{(i)} = \chi_i$ will not appear in (12). For an arbitrary function, however, its conjugate function is not given explicitly, and hence (12) is not an explicit optimization problem. As a result, to obtain an explicit formulation of the robust counterpart, one has to compute the conjugate functions of the constraint functions $g_j^{(i)}$, which except for very simple cases is not easy. This motivates us to investigate in the remainder of the paper under what conditions the robust counterpart in Theorem 3.2 can be further simplified, avoiding the computation of conjugate functions.

4. Explicit reformulation for robust counterparts. For any function f, let

$$\Re\mathcal{D}(f) = \bigcup_{x \in \mathrm{dom}(f)} \partial f(x);$$

that is, $\Re\mathcal{D}(f)$ is the range of the subdifferential mapping $\partial f(\cdot)$. If f is differentiable, $\Re\mathcal{D}(f)$ reduces to the range of its gradient mapping, i.e., $\Re\mathcal{D}(f) = \{\nabla f(x) : x \in \mathrm{dom}\ (f)\}$. In this section we make the following assumption.

Assumption 4.1. The functions $g_j^{(i)}(j = 1, \ldots, \ell^{(i)}, i = 1, \ldots, m)$ in (9) belong to the set of convex functions f that satisfy the condition

$$\mathrm{dom}(f^*) = \Re\mathcal{D}(f). \tag{17}$$

In fact, by the definition of subdifferential, the following relation always holds for any proper convex function: $\mathrm{dom}(f^*) \supseteq \Re\mathcal{D}(f)$. Condition (17) requires the converse also to be true. Indeed, condition (17) holds for many functions. It is evident that all convex functions defined on a subset of R^n with $\Re\mathcal{D}(f) = R^n$ satisfy condition (17). For example, when the function f is differentiable and strongly convex on R^n, the gradient $\nabla f(x)$ is a strongly monotone function from R^n to R^n. This implies that $\nabla f(x)$ is a bijective mapping [34, Theorem 6.4.4], and hence we have $\Re\mathcal{D}(f) = R^n$. A simple example is the quadratic function $f = \frac{1}{2}x^T Q x + bx + c$, where Q is a positive definite matrix; then $\Re\mathcal{D}(f) = \{Qx + b : x \in R^n\} = R^n$. When $\Re\mathcal{D}(f) \neq R^n$, (17) can still be satisfied in many cases. Later, we will show that all convex homogeneous of 1-degree functions satisfy (17) trivially, and $\Re\mathcal{D}(f)$ of any function of this class is a closed bounded region including the origin. Notice that for any (u, x) such that $u \in \partial f(x)$, we have $f^*(u) = u^T x - f(x)$. The importance of condition (17) is that under (17), for any $u \in \mathrm{dom}(f^*)$ there is $x \in \mathrm{dom}(f)$ such that $u \in \partial f(x)$ and therefore $f^*(u) = u^T x - f(x)$. Therefore, under Assumption 4.1, the robust counterpart

(12) can be represented explicitly. However, we omit the statement of this general result. We are now interested in functions that have more properties leading to further simplification of the robust counterpart.

We recall that a function $h : R^n \to \overline{R}$ is said to be positively homogeneous if there exists a constant $p > 0$ such that $h(\lambda x) = \lambda^p h(x)$ for all $\lambda \geq 0$ and $x \in \mathrm{dom}(h)$. If such a p exists, we simply say that the function h is homogeneous of p-degree. Notice that the definition implies $0 \in \mathrm{dom}(h)$ and $h(0) = 0$. We consider the linear space $\mathcal{L}_H$ generated by homogeneous functions; i.e., $\mathcal{L}_H$ is the collection of all functions that are finite linear combinations of homogeneous functions. Notice that for any real number α, $(\alpha h)(x)$ is also a homogeneous function if h is homogeneous. Therefore, $\mathcal{L}_H$ is the set of all finite sums of homogeneous functions. Clearly, a function f which is the sum of several homogeneous functions f_i is not necessarily homogeneous, unless all f_i have the same homogeneous degree. Linear space $\mathcal{L}_H$ includes many important classes of functions. Needless to say, all homogeneous functions (in particular, all norms $\|\cdot\|$) are in $\mathcal{L}_H$ and all polynomial functions are in $\mathcal{L}_H$.

The classical Euler homogeneous function theorem claims that if f is continuously differentiable and homogeneous of p-degree, then $pf(x) = x^T \nabla f(x)$, where $\nabla f(x)$ is the gradient of f. Below we establish a somewhat different version of the Euler homogeneous function theorem. This version allows the function to be nondifferentiable and nonhomogeneous but to belong to $\mathcal{L}_H$ and be convex.

LEMMA 4.1. *Let $f : R^n \to \overline{R}$ be a convex function in $\mathcal{L}_H$. Thus, f can be represented as $f(x) = f_1(x) + \cdots + f_N(x)$ for some N, where each f_i is homogeneous of p_i-degree, respectively.*

(i) *For any $x \in dom\ (f)$, we have*

$$\sum_{i=1}^{N} p_i f_i(x) = \inf_{y \in \partial f(x)} y^T x = \sup_{y \in \partial f(x)} y^T x;$$

i.e., for any $y \in \partial f(x)$, we have $\sum_{i=1}^{N} p_i f_i(x) = y^T x$.

(ii) *Suppose that $f : R^n \to \overline{R}$ is a convex function and is homogeneous of p-degree. Then for any $x \in \mathrm{dom}(f)$ and for any $y \in \partial f(x)$, we have $pf(x) = y^T x$.*

Proof. For any given $x \in \mathrm{dom}(f)$ and $y \in \partial f(x)$, by definition of subdifferential we have $f(u) \geq f(x) + y^T(u - x)$ for all $u \in \mathrm{dom}(f)$. Notice that $x \in \mathrm{dom}(f)$ if and only if $x \in \mathrm{dom}(f_i)$ for all $i = 1, \ldots, N$. Since all f_i's are homogeneous, for any $t > 0$, we have $u = tx \in \mathrm{dom}(f_i)$ for all $i = 1, \ldots, N$. This in turn implies that $u = tx \in \mathrm{dom}\ (f)$ for any $t > 0$. Setting $u = tx$ in the above inequality and by using homogeneity, we have

$$f(tx) = \sum_{i=1}^{N} f_i(tx) = \sum_{i=1}^{N} t^{p_i} f_i(x) \geq f(x) + y^T(tx - x)\ \forall\ t > 0,$$

i.e.,

$$\sum_{i=1}^{N} (t^{p_i} - 1) f_i(x) \geq (t - 1) y^T x\ \forall\ t > 0. \tag{18}$$

For $t > 1$, dividing both sides by $t-1$ and noting that y is any given element in $\partial f(x)$, we see from the above inequality that

$$\lim_{t \to 1_+} \sum_{i=1}^{N} \frac{t^{p_i} - 1}{t - 1} f_i(x) \geq \sup_{y \in \partial f(x)} y^T x.$$

Thus, we have $\sum_{i=1}^{N} p_i f_i(x) \geq \sup_{y \in \partial f(x)} y^T x$. Similarly, when $t < 1$, dividing both sides of (18) by $t-1$, we can prove that

$$\sum_{i=1}^{N} p_i f_i(x) = \lim_{t \to 1-} \sum_{i=1}^{N} \frac{t^{p_i} - 1}{t-1} f_i(x) \leq \inf_{y \in \partial f(x)} y^T x.$$

Combining the last two inequalities yields the desired result (i). Setting $N = 1$, we obtain the result (ii) from (i). ∎

Notice that when $N > 1$, Lemma 4.1 requires convexity of f but does not require convexity of individual functions f_i, which can be nonconvex. The next theorem is the main result of this section, which states that the robust counterpart can be represented explicitly by using only the nominal data and the constraint functions g_i together with their subdifferentials.

THEOREM 4.1. *Let K_i $(i = 1, \ldots, m)$ be given by (9), where each $g_j^{(i)} (j = 1, \ldots, \ell^{(i)}, i = 1, \ldots, m)$ is a closed proper convex function and belongs to the linear space $\mathcal{L}_H$, and is represented as*

$$g_j^{(i)}(x) = \sum_{k=1}^{m^{(ij)}} h_k^{(ij)}(x), \tag{19}$$

where each $h_k^{(ij)}(x)$ is homogeneous of $p_k^{(ij)}$-degree, and each $m^{(ij)} \geq 1$ is a given integer number. Let $g_j^{(i)}$ satisfy Assumption 4.1 and Slater's condition for each i. Then the robust programming problem (7) is equivalent to

$$\begin{aligned}
&\min c^T x \\
&s.t. \left(W^{(i)}(x)\right)^T \overline{M}^{(i)} V^{(i)}(x) + \sum_{j=1}^{\ell^{(i)}} \lambda_j^{(i)} \Delta_j^{(i)} + \Upsilon^{(i)} \leq b_i, \quad i = 1, \ldots, m, \\
&\chi_i = \begin{cases} \sum_{j \in J_i} u_j^{(i)} & \text{if } J_i \neq \emptyset, \\ 0 & \text{otherwise}, \end{cases} \quad i = 1, \ldots, m, \qquad (20) \\
&\lambda_j^{(i)} \geq 0, \quad j = 1, \ldots, \ell^{(i)}, \quad i = 1, \ldots, m, \\
&F(x) \leq 0,
\end{aligned}$$

where $\chi_i = W^{(i)}(x) \otimes V^{(i)}(x)$ and $J_i = \{j : \lambda_j^{(i)} > 0, j = 1, \ldots, \ell^{(i)}\}$, and

$$\Upsilon^{(i)} = \begin{cases} \sum_{j \in J_i} \lambda_j^{(i)} \left(\sum_{k=1}^{m^{(ij)}} (p_k^{(ij)} - 1) h_k^{(ij)}(w_j^{(i)}) \right) & \text{if } J_i \neq \emptyset, \\ 0 & \text{otherwise}, \end{cases}$$

where $w_j^{(i)}$ satisfies that $u_j^{(i)} / \lambda_j^{(i)} \in \partial g_j^{(i)}(w_j^{(i)})$ for $j \in J_i \neq \emptyset$.

Proof. Let $f \in \mathcal{L}_H$ be any convex function such that $f(x) = f_1(x) + \cdots + f_N(x)$, where f_i is homogeneous of p_i-degree, and let f satisfy condition (17). Let y^* be any element in $\mathrm{dom}(f^*) = \Re\mathcal{D}(f)$. This implies that there exists some point $x^* \in \mathrm{dom}\ (f)$ such that $y^* \in \partial f(x^*)$. Then, for any $x \in \mathrm{dom}(f)$, we have $f(x) \geq f(x^*) + (y^*)^T$

$(x-x^*)$, which can be written as $(y^*)^T x - f(x) \le (y^*)^T x^* - f(x^*)$ for all $x \in \mathrm{dom}(f)$. This, together with Lemma 4.1, implies that

$$f^*(y^*) = (y^*)^T x^* - f(x^*) = \sum_{i=1}^{N} p_i f_i(x^*) - f(x^*) = \sum_{i=1}^{N} (p_i - 1) f_i(x^*). \tag{21}$$

Setting $f = g_j^{(i)}$ and $y^* = u_j^{(i)}/\lambda_j^{(i)}$, where $g_j^{(i)}$ is given by (19), it follows from (21) that

$$\left(g_j^{(i)}\right)^* (u_j^{(i)}/\lambda_j^{(i)}) = \sum_{k=1}^{m^{(ij)}} (p_k^{(ij)} - 1) h_k^{(ij)}(w_j^{(i)}),$$

where $w_j^{(i)}$ can be any point such that $u_j^{(i)}/\lambda_j^{(i)} \in \partial g_j^{(i)}(w_j^{(i)})$. Substituting the above into Theorem 3.2, we have the desired result. □

We now consider the case in which all the functions $g_j^{(i)} (j = 1, \ldots, \ell^{(i)})$ are homogeneous. This is a special case of (19) with $m^{(ij)} = 1$ (for all $j = 1, \ldots, \ell^{(i)}, i = 1, \ldots, m$). We have the following result.

COROLLARY 4.1. *Let K_i be given by* (9)*, where each $g_j^{(i)} (j = 1, \ldots, \ell^{(i)})$ is convex and homogeneous of $p_j^{(i)}$-degree, and $g_j^{(i)}$ satisfy Assumption* 4.1*. Then the robust programming problem* (7) *is equivalent to* (20)*, but $\Upsilon^{(i)}$ is given as follows:*

$$\Upsilon^{(i)} = \begin{cases} \sum_{j \in J_i} (p_j^{(i)} - 1) \lambda_j^{(i)} g_j^{(i)}(w_j^{(i)}) & \text{if } J_i \neq \emptyset, \\ 0 & \text{otherwise,} \end{cases}$$

where $w_j^{(i)}$ satisfies that $u_j^{(i)}/\lambda_j^{(i)} \in \partial g_j^{(i)}(w_j^{(i)})$ for $j \in J_i \neq \emptyset$.

It is worth mentioning that $\Upsilon^{(i)}$ can be written as

$$\Upsilon^{(i)} = \begin{cases} \sum_{j \in J_i} \left(1 - 1/p_j^{(i)}\right) (u_j^{(i)})^T w_j^{(i)} & \text{if } J_i \neq \emptyset, \\ 0 & \text{otherwise.} \end{cases}$$

This follows from (ii) of Lemma 4.1. Actually, for any function f satisfying Assumption 4.1, (21) can also be written as $f^*(y^*) = (y^*)^T x^* - f(x^*) = (1 - 1/p)(y^*)^T x^*$. Therefore,

$$\left(g_j^{(i)}\right)^* (u_j^{(i)}/\lambda_j^{(i)}) = (1 - 1/p_j^{(i)})(u_j^{(i)})^T w_j^{(i)}/\lambda_j^{(i)}$$

for some $w_j^{(i)}$ such that $u_j^{(i)}/\lambda_j^{(i)} \in \partial g_j^{(i)}(w_j^{(i)})$.

Remark 4.1. (i) Notice that in Corollary 4.1 we do not require Slater's condition, since it was shown in [32] that for homogeneous convex optimization, Lagrangian duality results hold without Slater's condition.

(ii) It should be mentioned that Slater's condition in Theorem 4.1 is not essential, and can be removed in many situations, or enforced by slightly changing the constants $\Delta_j^{(i)}$ in (9). Any function g in the linear space $\mathcal{L}_H$ is the sum of some homogeneous functions whose value is zero at the origin. Thus $0 \in K_i$ implies that $0 = g_j^{(i)}(0) \le \Delta_j^{(i)}$ for $j = 1, \ldots, \ell^{(i)}$; i.e., all constants $\Delta_j^{(i)}$ must be nonnegative in (9) when $g_j^{(i)} \in \mathcal{L}_H$. If all $\Delta_j^{(i)}$ are positive, Slater's condition holds trivially (this is the situation in most

practical applications; for example, when $g_j^{(i)}$ is a norm, $\Delta_j^{(i)}$ is positive since otherwise the uncertainty set contains at most one point). If not all $\Delta_j^{(i)}$ are positive, replacing $\Delta_j^{(i)}$ in (9) by $\widehat{\Delta}_j^{(i)}$, where $\widehat{\Delta}_j^{(i)} = \Delta_j^{(i)}$ if $\Delta_j^{(i)} > 0$, and $\widehat{\Delta}_j^{(i)} = \varepsilon$ otherwise, for some small $\varepsilon > 0$, allows us to satisfy Slater's condition.

In the next section, we show that in homogeneous cases the above results can be further improved without making Assumption 4.1.

5. Homogeneous cases. We now show that for homogeneous of 1-degree functions, Assumption 4.1 holds trivially, and for a degree $p \neq 1$, a simple transformation will make the resulting functions satisfy Assumption 4.1. We also further simplify the reformulation. We first prove some basic properties of homogeneous functions. Part (i) of the following lemma in fact follows from [30], but for completeness we provide a simple proof. It appears that the result of part (ii) of the following lemma should be valid for nondifferentiable functions as well, but for simplicity of the proof we state it for twice differentiable functions.

LEMMA 5.1. *Let $f : \mathrm{dom}(f) \subseteq R^n \to R$ be convex and homogeneous of p-degree.*

(i) *If the degree $p > 1$, then $f(x) \geq 0$ over its domain, and if $p < 1$, then $f(x) \leq 0$ over its domain.*

(ii) *Let f be twice differentiable over its domain. Then for $p > 1$, the function $(f(x))^{1/p}$ is convex and homogeneous of 1-degree; for $p < 1$, the function $-(-f(x))^{1/p}$ is convex and homogeneous of 1-degree.*

Proof. Let x be any point in dom (f). By homogeneity and convexity of f, we have

$$(1/2)^p f(x) = f(x/2) \leq f(x)/2 + f(0)/2 = f(x)/2.$$

Thus, $[(1/2)^p - 1/2]\, f(x) \leq 0$, and hence the result (i) follows.

We now prove the result of part (ii). Consider the case of $p > 1$. By (i), $p > 1$ implies that $f(x) \geq 0$ over its domain. Let $\varepsilon > 0$ be any given positive number. Denote $g_\varepsilon(x) := (f(x) + \varepsilon)^{1/p}$. Notice that $\mathrm{dom}(g_\varepsilon) = \mathrm{dom}\ (f)$, and g_ϵ is twice differentiable. We prove first that g_ε is a convex function for any given $\varepsilon > 0$. It suffices to show that $\nabla^2 g_\varepsilon(x) \succeq 0$ (positive semidefinite). Since

$$\nabla^2 g_\varepsilon(x) = \frac{1}{p}(f(x) + \varepsilon)^{\frac{1}{p}-2} \left[\left(\frac{1}{p} - 1 \right) \nabla f(x) \nabla f(x)^T + (f(x) + \varepsilon) \nabla^2 f(x) \right],$$

it is sufficient to prove that

$$\left(\frac{1}{p} - 1 \right) \nabla f(x) \nabla f(x)^T + (f(x) + \varepsilon) \nabla^2 f(x) \succeq 0.$$

By Schur complementarity property, this is equivalent to showing that

$$\begin{bmatrix} \frac{p}{p-1}(f(x) + \varepsilon) & \nabla f(x)^T \\ \nabla f(x) & \nabla^2 f(x) \end{bmatrix} \succeq 0.$$

Thus, we need to show for all $(t, u) \in R^{n+1}$ that

$$\begin{aligned} \varphi(t,u) &= (t, u^T) \begin{bmatrix} \frac{p}{p-1}(f(x) + \varepsilon) & \nabla f(x)^T \\ \nabla f(x) & \nabla^2 f(x) \end{bmatrix} \begin{pmatrix} t \\ u \end{pmatrix} \\ &= \frac{p}{p-1} t^2 (f(x) + \varepsilon) + 2t \nabla f(x)^T u + u^T \nabla^2 f(x) u \geq 0. \end{aligned}$$

Case 1. $t = 0$. By convexity of f, $u^T \nabla^2 f(x) u \geq 0$ for any $u \in R^n$; thus we have $\varphi(t, u) \geq 0$.

Case 2. $t \neq 0$. In this case, it suffices to show that for any $u \in R^n$

$$\varphi(1, u) = \frac{p}{p-1}(f(x) + \varepsilon) + 2\nabla f(x)^T u + u^T \nabla^2 f(x) u \geq 0.$$

Since $\nabla^2 f(x) \succeq 0$, the function $\varphi(1, u)$ is convex with respect to u, and its minimum is attained if there exists some u^* such that

$$\nabla f(x) = -\nabla^2 f(x) u^*, \tag{22}$$

and the minimum value is

$$\varphi(1, u^*) = \frac{p}{p-1}(f(x) + \varepsilon) + \nabla f(x)^T u^*.$$

By Euler's formula, we have $x^T \nabla f(x) = p f(x)$. Differentiating both sides of this equation, we have $(p-1)\nabla f(x) = \nabla^2 f(x) x$, which shows that the vector $u^* = -\frac{1}{p-1} x$ satisfies (22); thus the minimum value is

$$\varphi(1, u^*) = \frac{p}{p-1}(f(x) + \varepsilon) - \frac{1}{p-1}\nabla f(x)^T x = \frac{p}{p-1}\varepsilon > 0.$$

The last equation follows from Euler's formula again. Therefore $\varphi(t, u) \geq 0$ for any $(t, u) \in R^{n+1}$. Convexity of $g_\varepsilon(x)$ follows. Since $\varepsilon > 0$ is arbitrary and $(f(x))^{1/p} = \lim_{\epsilon \to 0} g_\varepsilon(x)$, we conclude that $(f(x))^{1/p}$ is convex.

The case of $p < 1$ is considered analogously. □

According to our definition of a homogeneous function, its domain includes the origin. The next lemma shows that Assumption 4.1 is satisfied for any homogeneous of 1-degree convex function, and its subdifferential at the origin defines the domain of the conjugate function.

LEMMA 5.2. *Let $h : \mathrm{dom}(h) \subseteq R^N \to \overline{R}$ be a closed proper convex function and be homogeneous of* 1*-degree. Then*

$$\Re\mathcal{D}(h) = \bigcup_{x \in \mathrm{dom}(h)} \partial h(x) = \partial h(0).$$

Moreover, $\mathrm{dom}(h^*) = \Re\mathcal{D}(h) = \partial h(0)$.

Proof. Let z be any subgradient of h at x; then for any given y and any positive number λ we have $h(\lambda y) \geq h(x) + z^T(\lambda y - x)$. Since λ is positive, dividing both sides of the inequality by λ and using homogeneity of h, we have

$$h(y) \geq \frac{h(x) - z^T x}{\lambda} + z^T y.$$

Let $\lambda \to \infty$. We have $h(y) \geq z^T y$, which holds for any y. Consider the set

$$S := \{z : \ z^T y \leq h(y) \text{ for any } \ y \in \mathrm{dom}(h)\}.$$

From the above proof, we have seen that $\partial h(x) \subseteq S$ for any x, i.e., $\Re\mathcal{D}(h) \subseteq S$. In particular, we have $\partial h(0) \subseteq S$. Conversely, since $h(0) = 0$, we see that any $z \in S$ is a subgradient of h at $x = 0$. Thus, we have $S \subseteq \partial h(0)$. We conclude that $\Re\mathcal{D}(h) = S = \partial h(0)$. The first part of the lemma has been proved.

We now prove the second part of the lemma. For any $y^* \in \Re\mathcal{D}(h)$, there exists an x^* such that $y^* \in \partial f(x^*)$, and by definition of subgradient, we have that $(y^*)^T x - h(x) \leq (y^*)^T x^* - h(x^*)$ for any $x \in \mathrm{dom}(h)$, which implies that $h^*(y^*) < \infty$, i.e., $y^* \in \mathrm{dom}(h^*)$. Thus, the inclusion $\Re\mathcal{D}(h) \subseteq \mathrm{dom}(h^*)$ holds trivially (we mentioned this observation at the beginning of section 4).

Now we show that converse inclusion is also valid. Suppose that $y^* \in \mathrm{dom}(h^*)$. We show that $y^* \in S$. Notice that for homogeneous of 1-degree function h, $\mathrm{dom}(h)$ is a cone. Thus, for any given positive number λ, we have

$$\lambda h^*(y^*) = \sup_{x \in \mathrm{dom}(h)} (y^*)^T(\lambda x) - \lambda h(x) = \sup_{x \in \mathrm{dom}(h)} (y^*)^T(\lambda x) - h(\lambda x) = h^*(y^*).$$

Since $\lambda > 0$ can be any positive number, we have $h^*(y^*) = 0$, which in turn implies that $(y^*)^T x - h(x) \leq h^*(y^*) = 0$ for any $x \in \mathrm{dom}(h)$, and therefore $y^* \in S$. The desired result follows. ∎

We can now simplify the robust counterpart for the homogeneous 1-degree case.

THEOREM 5.1. *Let K_i be defined by* (9), *where all the functions $g_j^{(i)}$, $i = 1, \ldots, \ell^{(i)}$, are closed proper convex functions and are homogeneous of* 1*-degree. Then the robust counterpart* (7) *is equivalent to*

$$\begin{aligned}
&\min c^T x \\
&s.t.\ \left(W^{(i)}(x)\right)^T \overline{M}^{(i)} V^{(i)}(x) + \sum_{j=1}^{\ell^{(i)}} \lambda_j^{(i)} \Delta_j^{(i)} \leq b_i, \quad i = 1, \ldots, m, \\
&\chi_i = \begin{cases} \sum_{j \in J_i} u_j^{(i)} & \text{if } J_i \neq \emptyset, \\ 0 & \text{otherwise,} \end{cases} \quad i = 1, \ldots, m, \qquad (23) \\
&\lambda_j^{(i)} \geq 0, \quad j = 1, \ldots, \ell^{(i)}, \quad i = 1, \ldots, m, \\
&F(x) \leq 0,
\end{aligned}$$

where χ_i and J_i are the same as in Theorem 4.1, *and $u_j^{(i)}/\lambda_j^{(i)} \in \partial g_j^{(i)}(0)$ for $j \in J_i \neq \emptyset$, $i = 1, \ldots, m$.*

Proof. Under the conditions of the theorem, Lemma 5.2 claims that Assumption 4.1 holds, and, moreover, $\Re\mathcal{D}(g_j^{(i)}) = \partial g_j^{(i)}(0)$ for all $i = 1, \ldots, \ell^{(i)}$. From the proof of Theorem 4.1, when $J_i \neq \emptyset$, we can set $w_j^{(i)} = 0$, and hence $\Upsilon^{(i)} = (g_j^{(i)})^*(u_j^{(i)}/\lambda_j^{(i)}) = 0$. Thus, in this case, $\Upsilon^{(i)} \equiv 0$ no matter what J_i is. Therefore, the robust counterpart (7) eventually reduces to (23). As mentioned in Remark 4.1, we do not need Slater's condition for homogeneous cases. ∎

When $g_j^{(i)}$ is homogeneous of $p_j^{(i)}$-degree, where $p_j^{(i)} \neq 1$ and is twice differentiable, by (ii) of Lemma 5.1, we may transform it into a homogeneous of 1-degree function. Then we can use Theorem 5.1. When $p_j^{(i)} < 1$, by Lemma 5.1, the value of $g_j^{(i)}$ is nonpositive; thus the constraint $g_j^{(i)} \leq \Delta_j^{(i)}$ becomes redundant (since $\Delta_j^{(i)} \geq 0$) and thus can be removed from the list of constraints defining K_i. Therefore, without loss of generality, we assume that all $p_j^{(i)} \geq 1$. We now have the following result.

THEOREM 5.2. *Let K_i be defined by* (9), *where the functions $g_j^{(i)}$, $j = 1, \ldots, \ell^{(i)}$, are twice differentiable, convex, and homogeneous of $p_j^{(i)}$-degree ($p_j^{(i)} \geq 1$), respectively.*

Then the robust programming problem (7) *is equivalent to*

$$\min c^T x$$

$$s.t.\ \left(W^{(i)}(x)\right)^T \overline{M}^{(i)} V^{(i)}(x) + \sum_{j=1}^{\ell^{(i)}} \lambda_j^{(i)} \widetilde{\Delta_j^{(i)}} \le b_i, \quad i = 1, \ldots, m,$$

$$\chi_i = \begin{cases} \sum_{j \in J_i} u_j^{(i)} & \text{if } J_i \neq \emptyset, \\ 0 & \text{otherwise}, \end{cases} \quad i = 1, \ldots, m,$$

$$\lambda_j^{(i)} \ge 0, \quad j = 1, \ldots, \ell^{(i)}, \quad i = 1, \ldots, m,$$

$$F(x) \le 0,$$

where χ_i *and* J_i *are the same as in Theorem* 4.1, $u_j^{(i)} \in \lambda_j^{(i)} \partial \mathcal{G}_j^{(i)}(0)$ *for* $j \in J_i \neq \emptyset$, $i = 1, \ldots, m$, *and*

$$\mathcal{G}_j^{(i)} = \begin{cases} (g_j^{(i)})^{1/p_j^{(i)}}, & p_j^{(i)} > 1, \\ g_j^{(i)}, & p_j^{(i)} = 1, \end{cases} \quad \widetilde{\Delta_j^{(i)}} = \begin{cases} (\Delta_j^{(i)})^{1/p_j^{(i)}}, & p_j^{(i)} > 1, \\ \Delta_j^{(i)}, & p_j^{(i)} = 1. \end{cases} \tag{24}$$

Proof. We note that for $p_j^{(i)} > 1$, since $g_j^{(i)}$ and $\Delta_j^{(i)}$ are nonnegative by Lemma 5.1, the constraint $g_j^{(i)} \le \Delta_j^{(i)}$ in (9) is equivalent to $(g_j^{(i)})^{1/p_j^{(i)}} \le (\Delta_j^{(i)})^{1/p_j^{(i)}}$. Define $\mathcal{G}_j^{(i)}$ and $\widetilde{\Delta_j^{(i)}}$ as in (24). Then this result is an immediate consequence of Theorem 5.1 and Lemma 5.1. □

From Theorems 5.1 and 5.2, the structure of robust counterparts of uncertain optimization problems mainly depends on the subdifferentials of $g_j^{(i)}$ or $\mathcal{G}_j^{(i)}$ at the origin when functions $g_j^{(i)}$ are homogeneous.

Notice that any norm is convex and homogeneous of 1-degree and can be defined on the whole space. (But the converse is not true; for example, consider $f(t) : R \to R$ given by $f(t) = t$ if $t \ge 0$ and $f(t) = 2|t|$ if $t < 0$. Clearly, f is convex and homogeneous of 1-degree, but it is not a norm, because $f(-1) \neq f(1)$.) Theorem 5.1 can be immediately applied to the case of an uncertainty set defined by a finite system of norm inequalities. For this case, however, in addition to the above formulation of the robust counterpart via subgradients at the origin, we can further simplify it using dual norms and eliminating all variables $\lambda_j^{(i)}$. For any norm $\|\cdot\|$, we denote its dual norm by $\|\cdot\|_*$, i.e., $\|u\|_* = \sup_{\|x\| \le 1} u^T x$. When $g_j^{(i)}$ is a norm, we denote it by $\|\cdot\|^{(ij)}$ and its dual norm by $\|\cdot\|_*^{(ij)}$.

COROLLARY 5.1. *Let* K_i *be defined by* (9), *where all* $g_j^{(i)} (j = 1, \ldots, \ell^{(i)}, i = 1, \ldots, m)$ *are norms, denoted, respectively, by* $\|\cdot\|^{(ij)} (j = 1, \ldots, \ell^{(i)}, i = 1, \ldots, m)$;

then the robust counterpart (7) *is equivalent to*

$$\min c^T x$$

$$s.t.\ \left(W^{(i)}(x)\right)^T \overline{M}^{(i)} V^{(i)}(x) + \sum_{j=1}^{\ell^{(i)}} \Delta_j^{(i)} \left\| u_j^{(i)} \right\|_*^{(ij)} \le b_i, \quad i = 1, \ldots, m,$$

$$\chi_i = \sum_{j=1}^{\ell^{(i)}} u_j^{(i)}, \quad i = 1, \ldots, m,$$

$$F(x) \le 0,$$

where $\chi_i = W^{(i)}(x) \otimes V^{(i)}(x)$.

Proof. Notice that $u \in \partial\|0\|$ if and only if $u^T x \le \|x\|$ for any x which can be written as $u^T(x/\|x\|) \le 1$, i.e., $\|u\|_* \le 1$. Therefore, for $j \in J_i \ne \emptyset$, $u_j^{(i)}/\lambda_j^{(i)} \in \partial g_j^{(i)}(0)$ is equivalent to $\|\frac{u_j^{(i)}}{\lambda_j^{(i)}}\|_*^{(ij)} \le 1$, or just $\|u_j^{(i)}\|_*^{(ij)} \le \lambda_j^{(i)}$. Therefore, the constraints of (23) can be further written as

$$\left(W^{(i)}(x)\right)^T \overline{M}^{(i)} V^{(i)}(x) + \sum_{j=1}^{\ell^{(i)}} \lambda_j^{(i)} \Delta_j^{(i)} \le b_i, \quad i = 1, \ldots, m,$$

$$\chi_i = \begin{cases} \sum_{j \in J_i} u_j^{(i)} & \text{if } J_i \ne \emptyset, \\ 0 & \text{otherwise} \end{cases}, \quad i = 1, \ldots, m,$$

$$\lambda_j^{(i)} \ge 0, \quad j = 1, \ldots, \ell^{(i)}, \quad i = 1, \ldots, m,$$

$$\|u_j^{(i)}\|_*^{(ij)} \le \lambda_j^{(i)} \quad \forall j \in J_i \ne \emptyset, \quad i = 1, \ldots, m,$$

$$F(x) \le 0.$$

It is evident that the above system is equivalent to

$$\left(W^{(i)}(x)\right)^T \overline{M}^{(i)} V^{(i)}(x) + \sum_{j=1}^{\ell^{(i)}} \lambda_j^{(i)} \Delta_j^{(i)} \le b_i, \quad i = 1, \ldots, m,$$

$$\|u_j^{(i)}\|_*^{(ij)} \le \lambda_j^{(i)}, \quad j = 1, \ldots, \ell^{(i)}, \quad i = 1, \ldots, m,$$

$$\chi_i = \sum_{j=1}^{\ell^{(i)}} u_j^{(i)}, \quad i = 1, \ldots, m,$$

$$\lambda_j^{(i)} \ge 0, \quad j = 1, \ldots, \ell^{(i)}, \quad i = 1, \ldots, m,$$

$$F(x) \le 0.$$

Eliminating the variables $\lambda_j^{(i)}$, the above system becomes

$$\left(W^{(i)}(x)\right)^T \overline{M}^{(i)} V^{(i)}(x) + \sum_{j=1}^{\ell^{(i)}} \Delta_j^{(i)} \left\| u_j^{(i)} \right\|_*^{(ij)} \le b_i, \quad i = 1, \ldots, m,$$

$$\chi_i = \sum_{j=1}^{\ell^{(i)}} u_j^{(i)}, \quad i = 1, \ldots, m,$$

$$F(x) \le 0.$$

The desired result is obtained. □

6. Special cases. Complexity of robust counterparts depends both on the structure of the original optimization problems and on the structure of the uncertainty set. The harder the original optimization problem is and/or the more complex the uncertainty set is, the more difficult the robust counterpart is. In this section, we demonstrate how the general results developed above can be simplified by considering special optimization problems and/or special uncertainty sets. We take the LP problem as an example of a special optimization problem and take the widely used uncertainty set (10) as an example of a special uncertainty set. Thus we obtain new results for problem (1) with uncertainty set defined by (10) and for robust LP with general uncertainty sets. For this simplest of the considered cases (robust LP with uncertainty set (10)), we show that our results contain a number of related results in the literature, but they are under less restrictive assumptions, thus generalizing and strengthening these results.

6.1. Problem (1) with uncertainty set $\mathcal{U}$ defined by (10). Now we consider the uncertainty set (10), i.e.,

$$\mathcal{U} =: \left\{ D \;\middle|\; \exists z \in R^{|N|} : D = D_0 + \sum_{j \in N} \Delta D_j z_j, \quad \|z\| \le \Omega \right\}.$$

Since this model has been widely used in the literature (see, e.g., [5, 6, 7, 8, 9, 10, 11, 12, 13, 14]), it is interesting to see how our general results can be simplified when reduced to the above uncertainty set. Let H denote the matrix whose columns are $\Delta D_j, j = 1, \ldots, |N|$, i.e.,

$$H = [\Delta D_1, \ldots, \Delta D_{|N|}].$$

Define the function

$$g(u) = \inf\{\|z\| : Hz = u\}. \tag{25}$$

Then $g(u)$ is convex and homogeneous of 1-degree (convexity is proven in [36], and homogeneity can be checked directly). Now we show that the uncertainty set (10) can be represented equivalently in the form (9).

LEMMA 6.1. *Consider the uncertainty set $\mathcal{U}$ given by* (10). *Let $K = \{u|\ g(u) \le \Omega\}$, where g is given by* (25). *Then we have $K = \mathcal{U} - \{D_0\}$.*

Proof. Let u be any point in K. By the definition of $g(u)$, there exists a point z^* such that $g(u) = \|z^*\|$ and $Hz^* = u$. Since $u \in K$ implies $g(u) \le \Omega$, we have $\|z^*\| \le \Omega$. By the definition of $\mathcal{U}$, we see that $u \in \mathcal{U} - \{D_0\}$.

Conversely, suppose that $u \in \mathcal{U} - \{D_0\}$. Then there exists a point $D \in \mathcal{U}$ such that $u = D - D_0$. By the definition of $\mathcal{U}$, there exists a point z such that $u = Hz$ and $\|z\| \leq \Omega$. By the definition of g, this implies $g(u) \leq \Omega$, and hence $u \in K$. □

If the vectors $\{\Delta D_j : j = 1, \ldots, N\}$ are linearly independent, from $Hz = u$ we have $z = (H^T H)^{-1} H^T u$. Thus, we have

$$\mathcal{U} - \{D_0\} = K = \left\{ u \,\middle|\, g(u) = \|(H^T H)^{-1} H^T u\| \leq \Omega \right\}.$$

Since in general $|N|$ is less than the number of data of the problem, the term $H^T u$ can be zero even when $u \neq 0$. Thus, $g(u)$ is not a norm in this case, unless $\{\Delta D_j : j = 1, \ldots, N\}$ are linearly independent and $|N|$ equals the number of data of the problem, in which case H is an $|N| \times |N|$ invertible matrix.

Notice that K here has only one constraint which corresponds to the case $\ell^{(i)} = 1$ for all $i = 1, \ldots, m$, and by Theorem 16.3 in [36] the conjugate function of $g(u)$ is given by

$$g^*(w) = \begin{cases} 0 & \|H^T w\|_* \leq 1, \\ \infty & \text{otherwise}, \end{cases} \tag{26}$$

where $\|\cdot\|_*$ denotes the dual norm of $\|\cdot\|$.

We now consider our problem (1), where data $M^{(i)}$'s are subject to uncertainty of the type (10); i.e., for each i, the data $M^{(i)}$ belong to the set

$$\left\{ M^{(i)} \,\middle|\, \exists z \in R^{|N^{(i)}|} : M^{(i)} = \overline{M}_0^{(i)} + \sum_{j \in N^{(i)}} \Delta M_j^{(i)} z_j, \quad \|z\|^{(i)} \leq \Omega^{(i)} \right\}. \tag{27}$$

This can be written equivalently as

$$\mathcal{U}_i = \left\{ vec(M^{(i)}) \,\middle|\, \exists z \in R^{|N^{(i)}|} : vec(M^{(i)}) = vec(\overline{M}_0^{(i)}) + \sum_{j \in N^{(i)}} vec\left(\Delta M_j^{(i)}\right) z_j, \|z\|^{(i)} \leq \Omega^{(i)} \right\}, \tag{28}$$

$i = 1, \ldots, m$, where $N^{(i)}$ is the corresponding index set (not to be confused with N_i—the dimension of matrix $M^{(i)}$), and $\Omega^{(i)}$ is a given number. Note that we add the index (i) to the norm (i.e., $\|\cdot\|^{(i)}$), which allows us to use different norms for different constraints. Accordingly, we have the function

$$g^{(i)}(u) = \inf\{\|z\|^{(i)} : H^{(i)} z = u\},$$

where $H^{(i)} = [vec(\Delta M_1^{(i)}), vec(\Delta M_2^{(i)}), \ldots, vec(\Delta M_{|N^{(i)}|}^{(i)})]$, and thus by Lemma 6.1 we have

$$\mathcal{U}_i - \{vec(\overline{M}^{(i)})\} = K_i = \{u | g^{(i)}(u) \leq \Omega^{(i)}\}.$$

Using (26), we have

$$(g^{(i)})^*(w) = \begin{cases} 0, & \|(H^{(i)})^T w\|_*^{(i)} \leq 1, \\ \infty & \text{otherwise}. \end{cases} \tag{29}$$

Now we have all the necessary ingredients to develop our result. We first note that in this case, $\ell^{(i)} = 1$ for all $i = 1, \ldots, m$ since the uncertainty set $\mathcal{U}_i$ has only one

constraint $g^{(i)}(u) \le \Omega^{(i)}$. So, $\lambda^{(i)}$ is reduced to a scalar. Therefore, the constraints of the robust counterpart (12) that correspond to index i reduce to

$$\left(W^{(i)}(x)\right)^T \overline{M}^{(i)} V^{(i)}(x) + \lambda^{(i)}\Omega^{(i)} + \Upsilon^{(i)} \le b_i, \tag{30}$$

$$\chi_i = \begin{cases} u^{(i)}, & \lambda^{(i)} > 0, \\ 0, & \lambda^{(i)} = 0, \end{cases} \tag{31}$$

where

$$\Upsilon^{(i)} = \begin{cases} \lambda^{(i)}(g^{(i)})^*(u^{(i)}/\lambda^{(i)}), & \lambda^{(i)} > 0, \\ 0, & \lambda^{(i)} = 0. \end{cases} \tag{32}$$

When $\lambda^{(i)} > 0$, the system (30)–(32) becomes

$$\left(W^{(i)}(x)\right)^T \overline{M}^{(i)} V^{(i)}(x) + \lambda^{(i)}\Omega^{(i)} + \lambda^{(i)}(g^{(i)})^*(u^{(i)}/\lambda^{(i)}) \le b_i,$$

$$\chi_i = u^{(i)}.$$

Eliminating $u^{(i)}$ and using (29), the above system is equivalent to

$$\left(W^{(i)}(x)\right)^T \overline{M}^{(i)} V^{(i)}(x) + \lambda^{(i)}\Omega^{(i)} \le b_i,$$

$$\|(H^{(i)})^T(\chi_i)\|_*^{(i)} \le \lambda^{(i)}.$$

This can be written as

$$\left(W^{(i)}(x)\right)^T \overline{M}^{(i)} V^{(i)}(x) + \Omega^{(i)}\|(H^{(i)})^T\chi_i\|_*^{(i)} \le b_i. \tag{33}$$

When $\lambda^{(i)} = 0$, the system (30)–(32) is written as

$$\left(W^{(i)}(x)\right)^T \overline{M}^{(i)} V^{(i)}(x) \le b_i,$$

$$\chi_i = 0.$$

Clearly, this system can be written as (33), too. Hence, by Theorem 3.2, we have the following result.

THEOREM 6.1. *Under the uncertainty set* (27) *(or equally,* (28)*), the robust counterpart* (7) *is equivalent to*

$$\begin{aligned} &\min c^T x \\ &s.t.\ \left(W^{(i)}(x)\right)^T \overline{M}^{(i)} V^{(i)}(x) + \Omega^{(i)} \left\|(H^{(i)})^T\chi_i\right\|_*^{(i)} \le b_i, \quad i = 1, \ldots, m, \\ &\quad F(x) \le 0, \end{aligned}$$

where $\chi_i = W^{(i)}(x) \otimes V^{(i)}(x)$ *and* $H^{(i)} = [vec(\Delta M_1^{(i)}), vec(\Delta M_2^{(i)}), \ldots, \Delta vec(M_{N^{(i)}}^{(i)})]$.

6.2. LP with general uncertainty sets. Consider the LP problem discussed in section 2: $\min\{c^T x : \ Ax \le b, \ x \ge 0\}$, where $A \in R^{m\times n}, b \in R^m$, and $c \in R^n$. As discussed in section 2, without loss of generality, we assume that only the coefficients of A are subject to uncertainty.

There are two widely used ways to characterize the uncertain data of LP problems. One is the "row-wise" uncertainty model (a separate uncertainty set is specified for each row of A), and the other is what we may call the "global" uncertainty model (one uncertainty set for the whole matrix A is specified). We first consider the situation of "global" uncertainty.

Suppose that A is allowed to vary in such a way that its deviations from a given nominal $\overline{A}$ fall in a bounded convex set K of R^{mn} that contains the origin (zero). That is, the uncertainty set is defined as

$$\mathcal{U} = \{\widetilde{A} | vec(\widetilde{A}) - vec(\overline{A}) \in K\}, \tag{34}$$

where K is defined by convex inequalities:

$$K = \{u| \ g_j(u) \le \Delta_j, \ j = 1, \ldots, \ell\}. \tag{35}$$

Here Δ_j's are constants, and all g_j are closed proper convex functions. Then the robust counterpart of the LP problem with uncertainty set $\mathcal{U}$ is

$$\min\{c^T x : \widetilde{A}x \le b, \ x \ge 0 \ \forall \widetilde{A} \in \mathcal{U}\}. \tag{36}$$

First, from section 2 we know that for LP we can drop indexes i for $g_j^{(i)}$ and $\Delta_j^{(i)}$ in the previous discussion, since in the reformulation of LP as a special case of (1) and (2), the data matrix for each constraint (2) is the same; i.e., $M^{(i)} = \left[\begin{smallmatrix} A \\ 0 \end{smallmatrix}\right]_{n\times n}$ for all i (see section 2). Second, we note that for LP, the vector $\chi_i = W^{(i)} \otimes V^{(i)} = e_i \otimes x$ is linear in x. Therefore, the results in previous sections can be further simplified for LP. For example, Theorems 3.1, 3.2, and 5.2 and Corollary 5.1 can be stated as follows (Theorems 6.2 through 6.4 and Corollary 6.1, respectively).

THEOREM 6.2. *The robust LP problem* (36) *is equivalent to the convex programming problem*

$$\begin{aligned} &\min c^T x \\ &s.t. \ \bar{a}_i^T x + \delta^*(\chi_i | \mathrm{cl}(\mathrm{co}(K))) \le b_i, \quad i = 1, \ldots, m, \\ &\qquad x \ge 0, \end{aligned}$$

where $\mathrm{cl}(\mathrm{co}(K))$ *is the closed convex hull of the set* K, *and* $\chi_i = e_i \otimes x$.

Since $\delta^*(\cdot|\mathrm{cl}(\mathrm{co}K))$ is a closed convex function, the robust counterpart of any LP problem with the uncertainty set denoted by (34) and (35) is a convex programming problem.

THEOREM 6.3. *Let* K *be given by* (35), *where* $g_j (j = 1 \ldots, \ell)$, *are arbitrary closed proper convex functions. Suppose that Slater's condition holds; i.e., there exists a point* u_0 *such that* $g_j(u_0) < \Delta_j$ *for all* $j = 1, \ldots, \ell$. *Then the robust LP problem* (36)

is equivalent to

$$\min c^T x$$

$$s.t.\ \bar{a}_i^T x + \sum_{j=1}^{\ell} \lambda_j^{(i)} \Delta_j + \left(\sum_{j=1}^{\ell} \lambda_j^{(i)} g_j \right)^* (\chi_i) \le b_i, \quad i = 1, \ldots, m,$$

$$\lambda_j^{(i)} \ge 0, \quad j = 1, \ldots, \ell, \quad i = 1, \ldots, m,$$

$$x \ge 0,$$

or equivalently,

$$\min c^T x$$

$$s.t.\ \bar{a}_i^T x + \sum_{j=1}^{\ell} \lambda_j^{(i)} \Delta_j + \Upsilon^{(i)} \le b_i, \quad i = 1, \ldots, m,$$

$$\chi_i = \begin{cases} \sum_{j \in J_i} u_j^{(i)} & \text{if } J_i \ne \emptyset, \\ 0 & \text{otherwise,} \end{cases} \quad i = 1, \ldots, m, \tag{37}$$

$$\lambda_j^{(i)} \ge 0, \quad j = 1, \ldots, \ell, \quad i = 1, \ldots, m,$$

$$x \ge 0,$$

where $\chi_i = e_i \otimes x$, $J_i = \{j : \lambda_j^{(i)} > 0, j = 1, \ldots, \ell\}$, *and*

$$\Upsilon^{(i)} = \begin{cases} \sum_{j \in J_i} \lambda_j^{(i)} g_j^*(u_j^{(i)}/\lambda_j^{(i)}) & \text{if } J_i \ne \emptyset, \\ 0 & \text{otherwise.} \end{cases}$$

Remark 6.1. (i) For LP, the constraint "$\chi_i = \sum_{j \in J_i} u_j^{(i)}$" is a linear constraint.

(ii) It is well known that for any convex function f, the function $\hat{f}(x,t) = tf(x/t)$, where $t > 0$, is also convex in (x,t), and is positive homogeneous of 1-degree, that is, $\hat{f}(\alpha x, \alpha t) = \alpha \hat{f}(x,t)$, for any $\alpha > 0$. Problem (37) shows that all functions involved are homogeneous of 1-degree with respect to the variables $(x, \lambda^{(i)}, u^{(i)})$. Thus, the robust LP problem (36) is not only a convex programming problem but also a homogeneous programming problem, i.e., an optimization problem where all functions involved are homogeneous.

THEOREM 6.4. *Let K be defined by* (35), *where the functions $g_j, j = 1, \ldots, \ell$, are twice differentiable, convex, and homogeneous of p_j-degree ($p_j \ge 1$), respectively.*

Then the robust LP problem (36) *is equivalent to*

$$\begin{aligned}
&\min c^T x\\
&s.t.\ \bar{a}_i^T x + \sum_{j=1}^{\ell} \lambda_j^{(i)} \widetilde{\Delta_j} \le b_i, \quad i = 1, \dots, m,\\
&\chi_i = \begin{cases} \sum_{j \in J_i} u_j^{(i)} & \text{if } J_i \neq \emptyset, \\ 0 & \text{otherwise,} \end{cases} \qquad i = 1, \dots, m,\\
&\lambda_j^{(i)} \ge 0, \quad j = 1, \dots, \ell, \quad i = 1, \dots, m,\\
&x \ge 0,
\end{aligned}$$

where χ_i *and* J_i *are the same as in Theorem* 6.3, $u_j^{(i)} \in \lambda_j^{(i)} \partial \mathcal{G}_j(0)$ *for* $j \in J_i \neq \emptyset, i = 1, \dots, m$, *and*

$$\mathcal{G}_j = \begin{cases} (g_j)^{1/p_j}, & p_j > 1, \\ g_j, & p_j = 1, \end{cases} \qquad \widetilde{\Delta_j} = \begin{cases} (\Delta_j)^{1/p_j}, & p_j > 1, \\ \Delta_j, & p_j = 1. \end{cases}$$

COROLLARY 6.1. *Let* K *be defined by* (35), *where all* g_j $(j = 1, \dots, \ell)$ *are norms, denoted, respectively, by* $\|\cdot\|^{(j)}, j = 1, \dots, \ell$; *then the robust counterpart* (36) *is equivalent to*

$$\begin{aligned}
&\min c^T x\\
&s.t.\ \bar{a}_i^T x + \sum_{j=1}^{\ell} \Delta_j \|u_j^{(i)}\|_*^{(j)} \le b_i, \quad i = 1, \dots, m,\\
&e_i \otimes x = \sum_{j=1}^{\ell} u_j^{(i)}, \quad i = 1, \dots, m,\\
&x \ge 0.
\end{aligned}$$

Now we briefly discuss the situation of "row-wise" uncertainty sets. In this case, in order to apply our general results, we reformulate LP in the form (1) in a different way than in section 2. Consider functions $f_i(x)$ of the form (2), where $W^{(i)}(x) = e_i \in R^n$, $V^{(i)}(x) = x \in R^n$ (same as in section 2). Throughout the rest of the paper, we denote by $A_i (i = 1, \dots, m)$ the ith row of A. Thus, A_i is an n-dimensional row vector. The $n \times n$ matrix $M^{(i)}$ is the matrix having A_i as its ith row and 0 elsewhere, i.e.,

$$M^{(i)} = \begin{bmatrix} 0 \\ A_i \\ 0 \end{bmatrix}_{n \times n}, \quad i = 1, \dots, m. \tag{38}$$

Then the ith constraint of $Ax \le b$ can be written as

$$f_i = (W^{(i)})^T M^{(i)} V^{(i)} \le b_i \tag{39}$$

for $i = 1, \dots, m$. Then applying the results of sections 3, 4, and 5 to the optimization problem (1) with the above inequality constraints and $F(x) = -x \le 0$, we can obtain a formulation for robust LP with "row-wise" uncertainty sets. We omit these results.

The formulation for other special cases such as the LCP and QP can be derived similarly; we leave these derivations to interested readers.

6.3. LP with uncertainty set of type (10). In this section, we consider the LP problem $\min\{c^T x : Ax \le b, x \ge 0\}$ under uncertainty of type (10). We will show that our results in this section include a number of recent results on robust LP in the literature as special cases. From Theorems 6.1 and 6.3, we have the following result.

THEOREM 6.5. (i) *Under the "row-wise" uncertainty set*

$$\mathcal{U}_i = \left\{ A_i \,\middle|\, \exists u \in R^{N^{(i)}} : A_i = \overline{A}_i + \sum_{j \in N^{(i)}} \Delta A_j^{(i)} u_j,\ \|u\|^{(i)} \le \Omega^{(i)} \right\}, \tag{40}$$

the robust counterpart of LP is equivalent to

$$\begin{aligned} &\min c^T x \\ &s.t.\ \bar{a}_i^T x + \Omega^{(i)} \left\| \left(H^{(i)}\right)^T x \right\|_*^{(i)} \le b_i, \quad i = 1, \ldots, m, \\ &\quad x \ge 0, \end{aligned} \tag{41}$$

where the matrix $H^{(i)} = [(\Delta A_1^{(i)})^T, (\Delta A_2^{(i)})^T, \ldots, (\Delta A_{|N^{(i)}|}^{(i)})^T]$.

(ii) *Under the "global" uncertainty set*

$$\mathcal{U} = \left\{ A \,\middle|\, \exists u \in R^{|N|} : A = \overline{A} + \sum_{j \in N} \Delta A_j u_j,\ \|u\| \le \Omega \right\}, \tag{42}$$

where A is an $m \times n$ matrix, the robust counterpart of LP is equivalent to

$$\begin{aligned} &\min c^T x \\ &s.t.\ \bar{a}_i^T x + \Omega \left\| \widetilde{H}^T \widetilde{\chi}_i \right\|_* \le b_i, \quad i = 1, \ldots, m, \\ &\quad x \ge 0, \end{aligned} \tag{43}$$

where the matrix $\widetilde{H} = [vec(\Delta A_1), vec(\Delta A_2), \ldots, vec(\Delta A_{|N|})]$ *and* $\widetilde{\chi}_i = e_i^{(m)} \otimes x$, *where* $e_i^{(m)}$ *denotes the ith column of the $m \times m$ identity matrix. Equivalently, the inequality* (43) *can be written as*

$$\bar{a}_i^T x + \Omega \left\| \left(\widetilde{H}^{(i)}\right)^T x \right\|_* \le b_i, \quad i = 1, \ldots, m,$$

where the matrix $\widetilde{H}^{(i)} = [(\Delta A_1)^T e_i^{(m)}, (\Delta A_2)^T e_i^{(m)}, \ldots, (\Delta A_{|N|})^T e_i^{(m)}]$.

Proof. To prove the result (i), we show that it is an immediate corollary of Theorem 6.1. To apply Theorem 6.1, we first reformulate the LP problem in the form (1) as we did at the end of section 6.2. The ith constraint of $Ax \le b$, i.e., $A_i x \le b_i$, can be written as (39), where $M^{(i)}$ is given by (38). Clearly, we have

$$vec(M^{(i)}) = e_i \otimes A_i^T, \quad vec(\overline{M}^{(i)}) = e_i \otimes \overline{A}_i^T.$$

Notice that when A_i belongs to the uncertainty set (40), then the $vec(M^{(i)})$ belongs to the following uncertainty set:

$$\left\{ vec(M^{(i)}) \,\middle|\, \exists u \in R^{|N^{(i)}|} : vec(M^{(i)}) = e_i \otimes \overline{A}_i^T + \sum_{j \in N^{(i)}} \left(e_i \otimes (\Delta A_j^{(i)})^T\right) u_j,\ \|u\|^{(i)} \le \Omega^{(i)} \right\}.$$

By Theorem 6.1, robust LP is equivalent to

$$\begin{aligned} &\min c^T x \\ &\text{s.t. } \bar{a}_i^T x + \Omega^{(i)} \left\| \left(P^{(i)}\right)^T \chi_i \right\|_*^{(i)} \le b_i, \quad i = 1, \ldots, m, \\ &\qquad x \ge 0, \end{aligned}$$

where $\chi_i = e_i \otimes x$ and the matrix

$$P^{(i)} = \left[e_i \otimes (\Delta A_1^{(i)})^T,\ e_i \otimes (\Delta A_2^{(i)})^T, \ldots,\ e_i \otimes (\Delta A_{|N^{(i)}|}^{(i)})^T \right].$$

Notice that

$$\left(P^{(i)}\right)^T \chi_i = \left[(\Delta A_1^{(i)})^T, (\Delta A_2^{(i)})^T, \ldots, (\Delta A_{|N^{(i)}|}^{(i)})^T \right]^T x.$$

Therefore, the result (i) holds.

Using the uncertainty set (42), item (ii) can also be proved by applying Theorem 6.1. In fact, we can reformulate the LP problem in the form of (1) as in section 2, where all the data matrix $M^{(i)}$ are equal to $\left[\begin{smallmatrix} A \\ 0 \end{smallmatrix}\right]_{n \times n}$. Notice that the uncertainty set (42) can be written as

$$\left\{ vec\left(\left[\begin{smallmatrix} A \\ 0 \end{smallmatrix}\right]\right) \middle| \exists u \in R^{|N|} : vec\left(\left[\begin{smallmatrix} A \\ 0 \end{smallmatrix}\right]\right) = vec\left(\left[\begin{smallmatrix} \overline{A} \\ 0 \end{smallmatrix}\right]\right) + \sum_{j \in N} vec\left(\left[\begin{smallmatrix} \Delta A_j \\ 0 \end{smallmatrix}\right]\right) u_j,\ \|u\| \le \Omega \right\}.$$

This is the uncertainty set of the form (28). Thus, by Theorem 6.1, robust LP is equivalent to

$$\begin{aligned} &\min c^T x \\ &\text{s.t. } \bar{a}_i^T x + \Omega \|H^T \chi_i\|_* \le b_i, \quad i = 1, \ldots, m, \\ &\qquad x \ge 0, \end{aligned}$$

where $\chi_i = e_i \otimes x$ and the matrix

$$H = \left[vec\left(\left[\begin{smallmatrix} \Delta A_1 \\ 0 \end{smallmatrix}\right]\right), vec\left(\left[\begin{smallmatrix} \Delta A_2 \\ 0 \end{smallmatrix}\right]\right), \ldots, vec\left(\left[\begin{smallmatrix} \Delta A_{|N|} \\ 0 \end{smallmatrix}\right]\right) \right].$$

Denote by $\widetilde{\chi}_i = e_i^{(m)} \otimes x$, where $e_i^{(m)}$ denotes the ith column of the $m \times m$ identity matrix. It is easy to check that

$$H^T \chi_i = \widetilde{H}^T \widetilde{\chi}_i = \left(\widetilde{H}^{(i)}\right)^T x,$$

where the matrices

$$\widetilde{H} = \left[vec(\Delta A_1),\ vec(\Delta A_2),\ \ldots,\ vec(\Delta A_{|N|}) \right],$$

$$H^{(i)} = \left[(\Delta A_1)^T e_i^{(m)}, (\Delta A_2)^T e_i^{(m)}, \ldots, (\Delta A_{|N|})^T e_i^{(m)} \right].$$

Thus, the desired result (ii) follows. □

Notice that dual norms appear in (41) and (43). If the norms used are some special norms such as $\ell_1, \ell_2, \ell_\infty, \ell_1 \cap \ell_\infty, \ell_2 \cap \ell_\infty$, then their dual norms $\|\cdot\|_*$ are explicitly known (see, e.g., [14]).

In [12], Bertsimas, Pachamanova, and Sim studied the case of robust LP with uncertainty sets defined by general norms. Their result provides a unified treatment of the approaches in [23, 24, 6, 7, 11]. However, their result is a special case of Theorem 6.5. Their uncertainty set is defined by the inequality

$$\|M(vec(A) - vec(\overline{A}))\| \leq \Delta,$$

where M is an invertible matrix and Δ is a given constant. Clearly, this inequality can be written as

$$vec(A) = vec(\overline{A}) + M^{-1}u, \|u\| \leq \Delta.$$

This is a special case of the uncertainty model (42), corresponding to the case when $|N|$ is equal to the number of data and the perturbation directions ΔA_j's are linearly independent (here ΔA_j's are the column vectors of M^{-1}). So, when we apply Theorem 6.5(ii) to such a special uncertainty set, we obtain the same result as "Theorem 2" in [12]. But our result in Theorem 6.5(ii) is more general than the result in [12] because our result can even deal with the cases when the perturbation direction matrix H is singular and even not a square matrix.

It should be mentioned that "Theorem 2" in [12] can also be obtained from our Corollary 6.1. Since M is invertible, we can define the function $g(D) = \|MD\|$, which is a norm. The uncertainty set is defined by only one norm inequality, i.e., $g(D) \leq \Delta$. So, setting $\ell = 1$ in Corollary 6.1, we obtain "Theorem 2" in [12] again.

Now we compare Theorem 6.5 with the corresponding results for robust LP in Bertsimas and Sim [14]. For LP, Theorem 6.5(i) strengthens (generalizes) the corresponding result in [14] in the sense that we do not impose extra conditions on the norms, but in [14] a similar result is obtained under the additional assumption that the norms are absolute norms. Below we elaborate on this in more detail.

As we pointed out in section 2, without loss of generality, it is sufficient to consider the case when only A is subject to uncertainty. For LP, only "row-wise" uncertainty is considered in [14]; for the ith linear inequality $A_i x \leq b_i$, A_i belongs to the uncertainty set (40). Bertsimas and Sim [14] defined $f(x, A_i) = -(A_i x - b_i)$, and

$$s_j = g(x, \Delta A_j^{(i)}) =: \max\{-(\Delta A_j^{(i)})x, (\Delta A_j^{(i)})x\} = \left|(\Delta A_j^{(i)})x\right|, \quad j = 1, \ldots, N^{(i)}.$$

Bertsimas and Sim [14] proved that for LP, when the norm $\|\cdot\|^{(i)}$ used in (40) is an absolute norm, the robust LP constraint is equivalent to

$$f(x, \overline{A}_i) \geq \Omega^{(i)} \|s\|_*^{(i)} \quad \text{(or equally, } f(x, \overline{A}_i) \geq \Omega^{(i)} y,\ \|s\|_*^{(i)} \leq y).$$

That is,

$$-\overline{A}_i x - b_i \geq \Omega^{(i)} \left\| \left[(\Delta A_1^{(i)})^T, (\Delta A_2^{(i)})^T, \ldots, (\Delta A_{|N^{(i)}|}^{(i)})^T \right]^T x \right\|_*^{(i)},$$

which is

$$\overline{A}_i x + \Omega^{(i)} \left\| \left[(\Delta A_1^{(i)})^T, (\Delta A_2^{(i)})^T, \ldots, (\Delta A_{|N^{(i)}|}^{(i)})^T \right]^T x \right\|_*^{(i)} \leq b_i.$$

This is the same result as Theorem 6.5(i). So, Bertsimas and Sim [14] proved the result of Theorem 6.5(i) under the assumption that the norms used are absolute norms. We obtain this result without additional assumptions on the norms.

We can also apply our general results to nonlinear problems such as SOCP and QP. Let us comment on the differences of our approach from the approach of Bertsimas and Sim [14]. Applying our general results to robust QP would lead to *exact* formulations which, in general, would be computationally difficult. Bertsimas and Sim [14] aim at obtaining computationally tractable *approximate* formulations. These are two different ways of approaching nonlinear robust optimization problems. Computationally tractable approximate formulations are important for practical solution of large-scale problems: approximate solution is the price one has to pay for computational tractability. Exact formulations are also important. First, from a theoretical viewpoint, they allow us to gain more insight and to study the structure of the problems. Second, they can be used in practice to obtain exact solutions to small-scale problems. Third, they can provide new or strengthened results for important special cases when restricted to such cases, as demonstrated in this section.

7. Conclusion. One of our main goals was to show how the classic convex analysis tools can be used to study robust optimization. We showed that some rather general classes of robust optimization problems can be represented as explicit mathematical programming problems. We demonstrated how explicit reformulations of the robust counterpart of an uncertain optimization problem can be obtained if the uncertainty set is defined by convex functions that fall in the space $\mathcal{L}_H$ and satisfy the condition (17). Our strongest results correspond to the case where the functions defining the uncertainty set are homogeneous, because in this case the condition (17) holds trivially, and the robust counterpart can be further simplified. Our results provide a unified treatment of many situations that have been investigated in the literature. The analysis of this paper is applicable to much wider situations and more complicated uncertainty sets than those considered before; for example, it is applicable to cases where fluctuations of data may be asymmetric and not defined by norms.

REFERENCES

[1] I. Averbakh, *Minmax regret solutions for minimax optimization problems with uncertainty*, Oper. Res. Lett., 27 (2000), pp. 57–65.

[2] I. Averbakh, *On the complexity of a class of combinatorial optimization problems with uncertainty*, Math. Programming, 90 (2001), pp. 263–272.

[3] I. Averbakh, *Minmax regret linear resource allocation problems*, Oper. Res. Lett., 32 (2004), pp. 174–180.

[4] I. Averbakh and V. Lebedev, *Interval data minmax regret network optimization problems*, Discrete Appl. Math., 138 (2004), pp. 289–301.

[5] A. Ben-Tal, *Conic and Robust Optimization*, Lecture notes, University of Rome "La Sapienza," Rome, Italy, 2002.

[6] A. Ben-Tal and A. Nemirovski, *Robust convex optimization*, Math. Oper. Res., 23 (1998), pp. 769–805.

[7] A. Ben-Tal and A. Nemirovski, *Robust solutions to uncertain linear programs*, Oper. Res. Lett., 25 (1999), pp. 1–13.

[8] A. Ben-Tal and A. Nemirovski, *Robust solutions of linear programming problems contaminated with uncertain data*, Math. Programming, 88 (2000), pp. 411–424.

[9] A. Ben-Tal, A. Nemirovski, and C. Roos, *Robust solutions of uncertain quadratic and conic-quadratic problems*, SIAM J. Optim., 13 (2002), pp. 535–560.

[10] D. Bertsimas and M. Sim, *Robust discrete optimization and network flows*, Math. Programming, 98 (2003), pp. 49–71.

[11] D. Bertsimas and M. Sim, *The price of robustness*, Oper. Res., 52 (2004), pp. 35–53.

[12] D. Bertsimas, D. Pachamanova, and M. Sim, *Robust linear optimization under general norms*, Oper. Res. Lett., 32 (2004), pp. 510–516.

[13] D. Bertsimas and M. Sim, *Robust Discrete Optimization under Ellipsoidal Uncertainty Sets*, Report, MIT, Cambridge, MA, 2004.

[14] D. Bertsimas and M. Sim, *Tractable approximations to robust conic optimization problems*, Math Program., 107 (2006), pp. 5–36.
[15] D. Bienstock, *Experiments with robust optimization*, in Proceedings of the 19th International Symposium on Mathematical Programming, Rio de Janeiro, Brazil, 2006.
[16] Z. Chen and L. G. Epstein, *Ambiguity, risk, and asset returns in continuous time*, Econometrica, 70 (2002), pp. 1403–1443.
[17] J. Dupačová, *The minimax approach to stochastic program and illustrative application*, Stochastics, 20 (1987), pp. 73–88.
[18] J. Dupačová, *Stochastic programming: Minimax approach*, in Encyclopedia of Optimization, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001, pp. 327–330.
[19] L. G. Epstein and M. Schneider, *Recursive multiple-priors*, J. Econom. Theory, 113 (2003), pp. 1–31.
[20] L. G. Epstein and M. Schneider, *Learning under ambiguity*, Rev. Econom. Stud., 74 (2007), pp. 1275–1303.
[21] E. Erdoğan and G. Iyengar, *Ambiguous chance constrained problems and robust optimization*, Math. Program., 107 (2006), pp. 37–61.
[22] J. E. Falk, *Exact solutions of inexact linear programs*, Oper. Res., 24 (1976), pp. 783–787.
[23] L. El Ghaoui and H. Lebret, *Robust solutions to least-squares problems with uncertain data*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 1035–1064.
[24] L. El Ghaoui, F. Oustry, and H. Lebret, *Robust solutions to uncertain semidefinite programs*, SIAM J. Optim., 9 (1998), pp. 33–52.
[25] D. Goldfarb and G. Iyengar, *Robust convex quadratically constrained programs*, Math. Program., 97 (2003), pp. 495–515.
[26] D. Goldfarb and G. Iyengar, *Robust portfolio selection problems*, Math. Oper. Res., 28 (2003), pp. 1–38.
[27] L. P. Hansen and T. J. Sargent, *Robust control and model uncertainty*, Amer. Econ. Rev., 91 (2001), pp. 60–66.
[28] R. Jagannathan, *Minimax procedure for a class of linear programs under uncertainty*, Oper. Res., 25 (1977), pp. 173–177.
[29] P. Kouvelis and G. Yu, *Robust Discrete Optimization and Its Applications*, Kluwer Academic Publishers, Boston, 1997.
[30] G. LeBlanc, *Homogeneous programming: Saddlepoint and perturbation function*, Econometrica, 45 (1977), pp. 729–736.
[31] P. J. Maenhout, *Robust portfolio rules and asset pricing*, The Review of Financial Studies, 17 (2004), pp. 951–983.
[32] P. V. Moeseke, *Saddlepoint in homogeneous programming without Slater condition*, Econometrica, 42 (1974), pp. 593–596.
[33] J. M. Mulvey, R. J. Vanderbei, and S. A. Zenious, *Robust optimization of large scale systems*, Oper. Res., 43 (1995), pp. 264–281.
[34] J. M. Ortega and W. C. Rheinboldt, *Iterative Algorithms for Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.
[35] M. C. Pinar and R. H. Tütüncü, *Robust profit opportunities in risky financial portfolios*, Oper. Res. Lett., 33 (2005), pp. 331–340.
[36] R. T. Rockafellar, *Convex Analysis*, 2nd ed., Princeton University Press, Princeton, NJ, 1970.
[37] A. Shapiro and S. Ahmed, *On a class of minimax stochastic problems*, SIAM J. Optim., 14 (2004), pp. 1237–1249.
[38] A. Shapiro and A. J. Kleywegt, *Minimax analysis of stochastic problems*, Optim. Methods Softw., 17 (2002), pp. 523–542.
[39] A. L. Soyster, *A duality theory for convex programming with set-inclusive constraints*, Oper. Res., 22 (1974), pp. 892–898.
[40] A. L. Soyster, *Convex programming with set-inclusive constraints and applications to inexact linear programming*, Oper. Res., 21 (1973), pp. 1154–1157.
[41] Z. Wang, *A shrinkage approach to model uncertainty and asset allocation*, The Review of Financial Studies, 18 (2005), pp. 673–705.
[42] J. Žáčková, *On minimax solutions of stochastic linear programs*, Čas. Pęst. Mat., 91 (1966), pp. 423–430.
[43] Y. Zhang, *General robust-optimization formulation for nonlinear programming*, J. Optim. Theory Appl., 132 (2007), pp. 111–124.

SIAM J. OPTIM.
Vol. 18, No. 4, pp. 1467–1489

A METHOD OF CENTERS WITH APPROXIMATE SUBGRADIENT LINEARIZATIONS FOR NONSMOOTH CONVEX OPTIMIZATION*

KRZYSZTOF C. KIWIEL†

Abstract. We give a proximal bundle method for constrained convex optimization. It requires only evaluating the problem functions and their subgradients with an unknown accuracy ϵ. Employing a combination of the classic method of centers' improvement function with an exact penalty function, it does not need a feasible starting point. It asymptotically finds points with at least ϵ-optimal objective values that are ϵ-feasible. When applied to the solution of linear programming problems via column generation, it allows for ϵ-accurate solutions of column generation subproblems.

Key words. nondifferentiable optimization, convex programming, proximal bundle methods, approximate subgradients, column generation

AMS subject classifications. 65K05, 90C25

DOI. 10.1137/060668559

1. Introduction.

We are concerned with the solution of the following convex programming problem:

$$f_* := \inf\{\, f(u) : h(u) \le 0,\ u \in C \,\}, \tag{1.1}$$

where C is a "simple" closed convex set (typically a polyhedron) in the Euclidean space $\mathbb{R}^m$ with inner product $\langle\cdot,\cdot\rangle$ and norm $|\cdot|$, f and h are convex real-valued functions, and there exists a *Slater point*

$$\mathring{u} \in C \quad \text{such that} \quad h(\mathring{u}) < 0. \tag{1.2}$$

Further, we assume that for fixed (and possibly unknown) *accuracy tolerances* $\epsilon_f, \epsilon_h \ge 0$, for each $u \in C$ we can find *approximate values* f_u, h_u and *approximate subgradients* g_f^u, g_h^u that produce the *approximate linearizations* of f and h:

$$\bar{f}_u(\cdot) := f_u + \langle g_f^u, \cdot - u\rangle \le f(\cdot) \quad \text{with} \quad \bar{f}_u(u) = f_u \ge f(u) - \epsilon_f, \tag{1.3a}$$

$$\bar{h}_u(\cdot) := h_u + \langle g_h^u, \cdot - u\rangle \le h(\cdot) \quad \text{with} \quad \bar{h}_u(u) = h_u \ge h(u) - \epsilon_h. \tag{1.3b}$$

Thus $f_u \in [f(u) - \epsilon_f, f(u)]$ estimates $f(u)$, and $g_f^u \in \partial_{\epsilon_f} f(u)$; i.e., g_f^u is a member of

$$\partial_{\epsilon_f} f(u) := \{g : f(\cdot) \ge f(u) - \epsilon_f + \langle g, \cdot - u\rangle\},$$

the ϵ_f-subdifferential of f at u. Similar relations hold for f replaced by h.

This paper modifies the phase 1-phase 2 method of centers of [Kiw85, section 5.7] and extends it to approximate linearizations. We first discuss the *exact* case of $\epsilon_f = \epsilon_h = 0$. For an infeasible starting point, in phase 1 this method reduces the constraint violation while keeping the objective increase as small as possible; this is reasonable especially if the starting point is close to a solution. Once a feasible point is found, in phase 2 the method reduces the objective while maintaining feasibility. Both phases employ the same improvement function, and each iterate solves

*Received by the editors August 28, 2006; accepted for publication (in revised form) June 27, 2007; published electronically January 16, 2008.
http://www.siam.org/journals/siopt/18-4/66855.html
†Systems Research Institute, Polish Academy of Sciences, Newelska 6, 01–447 Warsaw, Poland (kiwiel@ibspan.waw.pl).

a subproblem with f and h approximated via accumulated linearizations, stabilized by a quadratic term centered at the best point found so far. For phase 1, the analysis of [Kiw85, section 5.7] established optimality of all cluster points of the iterates without discussing their existence. A nontrivial sufficient condition for their existence was recently given in [SaS05, Prop. 4.3(ii)] for a modified variant. We show that this condition may be expected to hold only if problem (1.1) has a Lagrange multiplier $\bar{\mu} \leq 1$ (cf. Remark 3.13(ii)). We extend this condition to $\bar{\mu} > 1$ by replacing the current objective value in the improvement function with the value of an exact penalty function for penalty parameters $\hat{c} \geq \bar{\mu} - 1$. In effect, our results (cf. Theorems 3.8, 3.9, and 3.12) extend the main convergence results of [Kiw85, Thm. 5.7.4] and [SaS05, Thms. 4.4–4.5]. It is crucial for large-scale implementations that our results hold for various aggregation schemes that control the size of each quadratic programming (QP) subproblem, including the schemes of [Kiw85, section 5.7] and [SaS05] (see Remark 4.1).

Our combination of improvement and penalty functions with suitable penalty parameter updates seems to be necessary for our extension to inexact evaluations (otherwise, the method could jam at phase 1 when the standard improvement function cannot be reduced by more than $\max\{\epsilon_f, \epsilon_h\}$ for the tolerances ϵ_f, ϵ_h of (1.3); see Remark 3.5). Our method generates iterates in the set C, having f-values of at most $f_* + \epsilon_f$ and h-values of at most ϵ_h asymptotically (cf. Theorems 3.8–3.10), without *any* additional boundedness assumptions (such as boundedness of the feasible set, or the sufficient conditions discussed above). In a sense, this is the strongest convergence result one could hope for. Our algorithmic constructions and analysis combine the inexact linearization framework of [Kiw06a] (in a simplified version that highlights its crucial ingredients; cf. [Kiw06b]) with fairly intricate properties of improvement and penalty functions which have not been used so far in bundle methods.

As for other bundle methods, we note that the exact penalty function methods of [Kiw87, Kiw91] require additionally that the set C be bounded and may converge slowly when their penalty parameter estimates are too high. The level methods of [LNN95] (also see [Kiw95, Fáb00, BTN05]) need boundedness of the set C as well. Similar boundedness assumptions are employed in the filter methods of [FlL99, KRSS07]. Except for [Fáb00], all these methods work with exact linearizations. The conic bundle variant of [KiL07] employs inexact linearizations and does not need artificial merit functions, but it requires the knowledge of a Slater point and f being "simple" (e.g., linear or quadratic). We show elsewhere how to handle inexact linearizations in an exact penalty method [Kiw07b] and a filter method [Kiw07a], the latter being based on the present paper.

Our work was partly motivated by possible applications in column generation approaches to integer programming problems [LüD05], which lead to linear programming (LP) problems with huge numbers of columns. When the dual LP problems can be formulated as (1.1) (cf. [BLM$^+$07, LüD05, Sav97]), our approach allows for ϵ_h-accurate solutions of column generation subproblems as well as for recovering approximate solutions to the primal problems. (See [Kiw05, KiL07] for related developments and numerical results.)

The paper is organized as follows. In section 2, after reviewing basic properties of penalty and improvement functions, we present our bundle method. Its convergence is analyzed in section 3. Several modifications are given in section 4. Applications to column generation for LP problems are studied in section 5.

2. The proximal bundle method of centers.

2.1. Lagrange multipliers and exact penalties. We first recall some basic duality results for problem (1.1) (cf. [Ber99, sections 5.1 and 5.3]).

Consider the *Lagrangian* $L(\cdot;\mu) := f(\cdot) + \mu h(\cdot)$ with $\mu \in \mathbb{R}$, the *dual function* $q(\mu) := \inf_C L(\cdot;\mu)$, and the *dual problem* $q_* := \sup_{\mathbb{R}_+} q$ of (1.1). Under our assumptions, $f_* = q_*$. If $f_* > -\infty$, the *dual optimal set* $M := \operatorname{Arg\,max}_{\mathbb{R}_+} q$ is nonempty and compact and consists of *Lagrange multipliers* $\mu \geq 0$ such that $q(\mu) = f_*$; if $f_* = -\infty$, $M := \emptyset$. Thus, the quantity $\bar{\mu} := \inf_{\mu \in M} \mu$ is the *minimal* Lagrange multiplier if $f_* > -\infty$, $\bar{\mu} = \infty$ otherwise.

For a *penalty parameter* $c \geq 0$, the *exact penalty function*

$$\pi(\cdot;c) := f(\cdot) + ch(\cdot)_+ \quad \text{with} \quad h(\cdot)_+ := \max\{h(\cdot), 0\} \tag{2.1}$$

satisfies $\inf_C \pi(\cdot;c) = f_* > -\infty$ iff $c \geq \bar{\mu}$ (cf. [Ber99, section 5.4.5]).

2.2. Improvement functions. We associate with problem (1.1) the *improvement functions* defined for $\tau \in \mathbb{R}$ by

$$e(\cdot;\tau) := \max\{f(\cdot) - \tau, h(\cdot)\}, \quad e_C(\cdot;\tau) := e(\cdot;\tau) + i_C(\cdot), \quad E(\tau) := \inf e_C(\cdot;\tau), \tag{2.2}$$

where i_C is the *indicator* function of C ($i_C(u) = 0$ if $u \in C$, ∞ if $u \notin C$). In our context, τ will be an asymptotic estimate of f_* generated by our method, and to prove that $\tau \leq f_*$, we shall need the main property of the function E given in part (vi) of the lemma below.

LEMMA 2.1. (i) *The function E defined by* (2.2) *is nonincreasing and convex.*

(ii) *If E is improper, then $E(\cdot) = f_* = -\infty$ for f_* given by* (1.1).

(iii) *If E is proper, then E is Lipschitzian with modulus* 1.

(iv) *If E is proper and $f_* = -\infty$, then $E(\cdot) = \inf_C h \in (-\infty, 0)$.*

(v) *If $f_* > -\infty$, then $E(\tau) > 0$ for $\tau < f_*$, $E(f_*) = 0$, and $E(\tau) < 0$ for $f_* < \tau$.*

(vi) *If $E(\tau) \geq 0$ for some $\tau \in \mathbb{R}$, then $\tau \leq f_*$.*

Proof. (i) Monotonicity is obvious, and convexity follows from [Roc70, Thm. 5.7].

(ii) Since $\operatorname{dom} E = \mathbb{R}$, we have $E(\cdot) = -\infty$ by [Roc70, Thm. 7.2], and then $f_* = -\infty$ by (1.1).

(iii) E is finite on $\operatorname{dom} E = \mathbb{R}$, and $e(\cdot;\tau') \leq e(\cdot;\tau) + |\tau - \tau'|$ for any τ and τ'.

(iv) Since $f_* = -\infty$ implies $E(\cdot) \leq 0$, $E(\cdot)$ is constant and finite by [Roc70, Cor. 8.6.2], i.e., $E(\cdot) = \alpha \in \mathbb{R}$. Then, on the one hand, $\alpha \geq \inf_C h$ by (2.2). On the other hand, for $u \in C$ and $\tau \geq f(u) - h(u)$, the fact that $e(u;\tau) \leq h(u)$ yields $\alpha \leq \inf_C h < 0$ by (1.2).

(v) We have $E(f_*) \leq 0$ by (1.1), and $E(f_*) \geq 0$ (otherwise $f(u) < f_*$ and $h(u) < 0$ for some $u \in C$ would contradict (1.1)); thus $E(f_*) = 0$. By (1.2), for $\mathring{\tau} := f(\mathring{u}) - h(\mathring{u}) > f(\mathring{u}) \geq f_*$, $e(\mathring{u};\mathring{\tau}) = h(\mathring{u}) < 0$ implies $E(\mathring{\tau}) < 0$; so by convexity (consider the secant line $\bar{E}(\tau) := E(\mathring{\tau})(\tau - f_*)/(\mathring{\tau} - f_*)$), we have $E(\tau) > 0$ for $\tau < f_*$, $E(\tau) < 0$ for $\tau \in (f_*, \mathring{\tau}]$, and $E(\tau) < 0$ for $\tau > \mathring{\tau}$ by monotonicity.

(vi) E is proper by (ii), $f_* > -\infty$ by (iv), and (v) yields the conclusion. ∎

Let $U := \{u \in C : h(u) \leq 0\}$ and $U_* := \operatorname{Arg\,min}_U f$ denote the *feasible* and *optimal* sets of problem (1.1). We shall need the following extension of [Kiw85, Lem. 1.2.16].

LEMMA 2.2. *Let $\bar{u} \in C$, $\bar{c} \geq 0$, $\bar{\tau} := \pi(\bar{u};\bar{c})$ (cf.* (2.1)). *Then the following are equivalent*:

(a) $\bar{u} \in U_*$ (*i.e., $\bar{u}$ solves problem* (1.1));

(b) $E(\bar\tau) = e_C(\bar u;\bar\tau)$ (*i.e., $\bar u$ minimizes $e(\cdot;\bar\tau)$ over C*);
(c) $0 \in \partial e_C(\bar u;\bar\tau)$ (*i.e., $0 \in \partial\psi(\bar u)$, where $\psi(\cdot) := e_C(\cdot;\bar\tau)$*).

Proof. First, (a) implies $\bar\tau = f(\bar u) = f_*$, $e(\bar u;\bar\tau) = 0$, $E(\bar\tau) = 0$ by Lemma 2.1(v), and hence (b). Since (b) means $\bar u \in \operatorname{Arg\,min} e_C(\cdot;\bar\tau)$, (b) and (c) are equivalent. Next, note that

$$\partial e_C(\bar u;\bar\tau) = \partial i_C(\bar u) + \begin{cases} \partial f(\bar u) & \text{if} \quad f(\bar u) - \bar\tau > h(\bar u), \\ \operatorname{co}\{\partial f(\bar u) \cup \partial h(\bar u)\} & \text{if} \quad f(\bar u) - \bar\tau = h(\bar u), \\ \partial h(\bar u) & \text{if} \quad f(\bar u) - \bar\tau < h(\bar u). \end{cases} \tag{2.3}$$

Finally, (c) implies $h(\bar u) \le 0$ (otherwise $h(\bar u) > 0 \ge f(\bar u) - \bar\tau$ and $0 \in \partial e_C(\bar u;\bar\tau) = \partial h(\bar u) + \partial i_C(\bar u)$ would give $\min_C h = h(\bar u) > 0$, contradicting (1.2)); so the facts that $\bar\tau = f(\bar u)$ and $E(\bar\tau) = e(\bar u;\bar\tau) = 0$ yield $\bar\tau = f_*$ by Lemma 2.1(v), and hence (a). ∎

Lemma 2.2 suggests the following algorithmic scheme: Given the current iterate $\hat u \in C$ and the target $\hat\tau := \pi(\hat u;\hat c)$ for a penalty parameter $\hat c \ge 0$, find an approximate minimizer u of $e_C(\cdot;\hat\tau)$, replace $\hat u$ by u, and repeat. Note that if $e_C(u;\hat\tau) < e_C(\hat u;\hat\tau)$, then u is better than $\hat u$: either $f(u) < f(\hat u)$ and $u \in U$ if $\hat u \in U$, or $h(u) < h(\hat u)$ if $\hat u \notin U$. To progress towards the optimal set U_*, it helps if $e_C(\bar u;\hat\tau) \le e_C(\hat u;\hat\tau)$ for any optimal $\bar u \in U_*$; the sufficient condition given below employs the minimal multiplier $\bar\mu$ of section 2.1.

LEMMA 2.3. *Let $\bar u \in U_*$, $\hat u \in C$, $\hat c \ge 0$, $\hat\tau := \pi(\hat u;\hat c)$. Then $e(\hat u;\hat\tau) = h(\hat u)_+$, and $e(\bar u;\hat\tau) \le e(\hat u;\hat\tau)$ iff $f(\bar u) \le \pi(\hat u;\hat c+1)$. In particular, $f(\bar u) \le \pi(\hat u;\hat c+1)$ if $\hat c \ge \bar\mu - 1$.*

Proof. First, $\hat\tau = f(\hat u)$ and $e(\hat u;\hat\tau) = 0$ if $h(\hat u) \le 0$, $e(\hat u;\hat\tau) = h(\hat u)$ if $h(\hat u) > 0$. Next,

$$e(\bar u;\hat\tau) - e(\hat u;\hat\tau) = \max\{f(\bar u) - \pi(\hat u;\hat c+1), h(\bar u) - h(\hat u)_+\}$$

is nonpositive iff $f_* = f(\bar u) \le \pi(\hat u;\hat c+1)$; the latter holds if $\hat c + 1 \ge \bar\mu$ (see section 2.1). ∎

2.3. An overview of the method. Our method generates a sequence of *trial points* $\{u^k\}_{k=1}^\infty \subset C$ for evaluating the approximate values $f_u^k := f_{u^k}$, $h_u^k := h_{u^k}$, subgradients $g_f^k := g_f^{u^k}$, $g_h^k := g_h^{u^k}$, and linearizations $f_k := \bar f_{u^k}$, $h_k := \bar h_{u^k}$ of f and h at u^k, respectively, such that

$$f_k(\cdot) = f_u^k + \langle g_f^k, \cdot - u^k\rangle \le f(\cdot) \quad \text{with} \quad f_k(u^k) = f_u^k \ge f(u^k) - \epsilon_f, \tag{2.4a}$$

$$h_k(\cdot) = h_u^k + \langle g_h^k, \cdot - u^k\rangle \le h(\cdot) \quad \text{with} \quad h_k(u^k) = h_u^k \ge h(u^k) - \epsilon_h, \tag{2.4b}$$

as stipulated in (1.3). At iteration k, the polyhedral *cutting-plane models* of f and h

$$\check f_k(\cdot) := \max_{j\in J_f^k} f_j(\cdot) \le f(\cdot) \quad \text{with} \quad k \in J_f^k \subset \{1,\dots,k\}, \tag{2.5a}$$

$$\check h_k(\cdot) := \max_{j\in J_h^k} h_j(\cdot) \le h(\cdot) \quad \text{with} \quad k \in J_h^k \subset \{1,\dots,k\}, \tag{2.5b}$$

which stem from the accumulated linearizations, yield the relaxed version of problem (1.1)

$$\check f_*^k := \inf\big\{\check f_k(u) : u \in \check H_k \cap C\big\} \quad \text{with} \quad \check H_k := \big\{u : \check h_k(u) \le 0\big\}, \tag{2.6}$$

in which $\check{H}_k$ is an outer approximation of $H := \{u : h(u) \le 0\}$. The current *prox* (or *stability*) *center* $\hat{u}^k := u^{k(l)} \in C$ for some $k(l) \le k$ has the values $f^k_{\hat{u}} = f^{k(l)}_u$ and $h^k_{\hat{u}} = h^{k(l)}_u$:

$$f^k_{\hat{u}} \in [f(\hat{u}^k) - \epsilon_f, f(\hat{u}^k)] \quad \text{and} \quad h^k_{\hat{u}} \in [h(\hat{u}^k) - \epsilon_h, h(\hat{u}^k)]. \tag{2.7}$$

As in (2.2) and Lemma 2.2, our improvement function for subproblem (2.6) is given by

$$\check{e}_k(\cdot) := \max\{\check{f}_k(\cdot) - \tau_k, \check{h}_k(\cdot)\} \quad \text{with} \quad \tau_k := f^k_{\hat{u}} + c_k[h^k_{\hat{u}}]_+ \tag{2.8}$$

for some penalty coefficient $c_k \ge 0$ and $[\cdot]_+ := \max\{\cdot, 0\}$. We solve a proximal version of the relaxed improvement problem $\check{E}_k := \inf \check{e}^k_C$ with $\check{e}^k_C := \check{e}_k + i_C$ by finding the trial point

$$u^{k+1} := \arg\min \left\{ \phi_k(\cdot) := \check{e}_k(\cdot) + i_C(\cdot) + \tfrac{1}{2t_k}|\cdot - \hat{u}^k|^2 \right\}, \tag{2.9}$$

where $t_k > 0$ is a *stepsize* that controls the size of $|u^{k+1} - \hat{u}^k|$. For deciding whether u^{k+1} is better than $\hat{u}^k$, we use approximate values of the improvement function $e(\cdot; \tau_k)$. Thus, $e(\hat{u}^k; \tau_k)$ is approximated by $[h^k_{\hat{u}}]_+$, and $e(\hat{u}^k; \tau_k) - \check{e}_k(u^{k+1})$ by the *predicted decrease*

$$v_k := [h^k_{\hat{u}}]_+ - \check{e}_k(u^{k+1}). \tag{2.10}$$

When $f^k_{\hat{u}} < \check{f}_k(\hat{u}^k)$ or $h^k_{\hat{u}} < \check{h}_k(\hat{u}^k)$ due to inexact evaluations, v_k may be nonpositive; if necessary, we increase t_k, as well as c_k in (2.8) if $h^k_{\hat{u}} > 0$, and recompute u^{k+1} to decrease $\check{e}_k(u^{k+1})$ until $v_k \ge |u^{k+1} - \hat{u}^k|^2/2t_k$ (as motivated below). Of course, $e(u^{k+1}; \tau_k)$ is approximated by $\max\{f^{k+1}_u - \tau_k, h^{k+1}_u\}$. A *descent* step to $\hat{u}^{k+1} := u^{k+1}$ occurs if $\max\{f^{k+1}_u - \tau_k, h^{k+1}_u\} \le [h^k_{\hat{u}}]_+ - \kappa v_k$ for a fixed $\kappa \in (0, 1)$. Otherwise, a *null* step $\hat{u}^{k+1} := \hat{u}^k$ improves the next models $\check{f}_{k+1}$, $\check{h}_{k+1}$ with the new linearizations f_{k+1} and h_{k+1} (cf. (2.5)).

2.4. Aggregate linearizations and an optimality estimate. Extending the approach of [Kiw06a], we now use optimality conditions for subproblem (2.9) to derive aggregate linearizations (i.e., affine minorants) of the problem functions at u^{k+1} as well as an optimality estimate (see (2.22) below) related to Lemma 2.1(vi).

LEMMA 2.4. (i) *There exist subgradients* p^k_f, p^k_h, p^k_C *and a multiplier* ν_k *such that*

$$p^k_f \in \partial \check{f}_k(u^{k+1}),\ p^k_h \in \partial \check{h}_k(u^{k+1}),\ p^k_C \in \partial i_C(u^{k+1}), \tag{2.11}$$

$$\nu_k p^k_f + (1 - \nu_k)p^k_h + p^k_C = -\big(u^{k+1} - \hat{u}^k\big)/t_k, \tag{2.12}$$

$$\nu_k \in [0, 1],\ \nu_k[\check{e}_k(u^{k+1}) - \check{f}_k(u^{k+1}) + \tau_k] = 0,\ (1 - \nu_k)[\check{e}_k(u^{k+1}) - \check{h}_k(u^{k+1})] = 0. \tag{2.13}$$

(ii) *These subgradients determine the following* aggregate linearizations*:*

$$\bar{f}_k(\cdot) := \check{f}_k(u^{k+1}) + \langle p^k_f, \cdot - u^{k+1}\rangle \le \check{f}_k(\cdot) \le f(\cdot), \tag{2.14}$$

$$\bar{h}_k(\cdot) := \check{h}_k(u^{k+1}) + \langle p^k_h, \cdot - u^{k+1}\rangle \le \check{h}_k(\cdot) \le h(\cdot), \tag{2.15}$$

$$\bar{i}^k_C(\cdot) := i_C(u^{k+1}) + \langle p^k_C, \cdot - u^{k+1}\rangle \le i_C(\cdot), \tag{2.16}$$

$$\bar{e}^k_C(\cdot) := \nu_k[\bar{f}_k(\cdot) - \tau_k] + (1 - \nu_k)\bar{h}_k(\cdot) + \bar{i}^k_C(\cdot) \le \check{e}^k_C(\cdot) \le e_C(\cdot; \tau_k). \tag{2.17}$$

(iii) *For the* aggregate subgradient *and the* aggregate linearization error *given by*

$$p^k := \nu_k p_f^k + (1-\nu_k)p_h^k + p_C^k = \big(\hat{u}^k - u^{k+1}\big)/t_k \quad \textit{and} \quad \epsilon_k := [h_{\hat{u}}^k]_+ - \bar{e}_C^k(\hat{u}^k) \tag{2.18}$$

and the optimality measure

$$V_k := \max\big\{|p^k|, \epsilon_k + \langle p^k, \hat{u}^k\rangle\big\}, \tag{2.19}$$

we have

$$\bar{e}_C^k(\cdot) = \check{e}_k(u^{k+1}) + \langle p^k, \cdot - u^{k+1}\rangle, \tag{2.20}$$

$$[h_{\hat{u}}^k]_+ - \epsilon_k + \langle p^k, \cdot - \hat{u}^k\rangle = \bar{e}_C^k(\cdot) \le \check{e}_C^k(\cdot) \le e_C(\cdot;\tau_k), \tag{2.21}$$

$$e_C(u;\tau_k) \ge \check{e}_C^k(u) \ge \big[h_{\hat{u}}^k\big]_+ - V_k\big(1+|u|\big) \quad \textit{for all } u. \tag{2.22}$$

Proof. (i) Use the optimality condition $0 \in \partial\phi_k(u^{k+1})$ for (2.9) and the form (2.8) of $\check{e}_k$.

(ii) The first inequalities in (2.14)–(2.15) stem from (2.11) and the final ones from (2.5). Similarly, (2.11) gives (2.16) with $i_C(u^{k+1}) = 0$. Then (2.17) follows from the facts that $\nu \in [0,1]$ (cf. (2.13)) yields $\nu_k(\bar{f}_k - \tau_k) + (1-\nu_k)\bar{h}_k \le \check{e}_k$ by using $\bar{f}_k \le \check{f}_k$ and $\bar{h}_k \le \check{h}_k$ in (2.8) and that $\check{e}_C^k := \check{e}_k + i_C \le e_C(\cdot;\tau_k)$ by using $\check{f}_k \le f$ and $\check{h}_k \le h$ in (2.2).

(iii) For (2.20), use (2.12)–(2.13) and the definitions in (2.14)–(2.18); since $\bar{e}_C^k$ is affine, its expression in (2.21) follows from (2.18). Finally, since by the Cauchy–Schwarz inequality,

$$-\langle p^k, u\rangle + \epsilon_k + \langle p^k, \hat{u}^k\rangle \le |p^k||u| + \epsilon_k + \langle p^k, \hat{u}^k\rangle \le \max\{|p^k|, \epsilon_k + \langle p^k, \hat{u}^k\rangle\}(1+|u|)$$

in (2.21), we obtain (2.22) from the definition of V_k in (2.19). ∎

Observe that V_k is an optimality measure at phase 2: if $V_k = 0$ in (2.22), then $E(\tau_k) \ge 0$ gives $f_{\hat{u}}^k \le \tau_k \le f_*$ by Lemma 2.1(vi); similar relations hold asymptotically.

2.5. Ensuring sufficient predicted decrease. In view of the optimality estimate (2.22), we would like V_k to vanish asymptotically. Hence it is crucial to bound V_k via the predicted decrease v_k, since normally bundling and descent steps drive v_k to 0. The necessary bounds are given below.

LEMMA 2.5. (i) *In the notation of* (2.18)*, the predicted decrease* v_k *of* (2.10) *satisfies*

$$v_k = t_k|p^k|^2 + \epsilon_k. \tag{2.23}$$

(ii) *We have* $v_k \ge -\epsilon_k \Leftrightarrow t_k|p^k|^2/2 \ge -\epsilon_k \Leftrightarrow v_k \ge t_k|p^k|^2/2 = |u^{k+1} - \hat{u}^k|/2t_k$.

(iii) *For the* maximal evaluation error $\epsilon_{\max} := \max\{\epsilon_f, \epsilon_h\}$*, we have*

$$-\epsilon_k \le \epsilon_{\max}. \tag{2.24}$$

(iv) *The optimality measure of* (2.19) *satisfies* $V_k \le \max\{|p^k|, \epsilon_k\}(1+|\hat{u}^k|)$*. Moreover,*

$$v_k \ge \max\big\{t_k|p^k|^2/2, |\epsilon_k|\big\} \qquad \textit{if} \quad v_k \ge -\epsilon_k, \tag{2.25}$$

$$V_k \le \max\big\{(2v_k/t_k)^{1/2}, v_k\big\}\big(1+|\hat{u}^k|\big) \quad \textit{if} \quad v_k \ge -\epsilon_k, \tag{2.26}$$

$$V_k < (2\epsilon_{\max}/t_k)^{1/2}\big(1+|\hat{u}^k|\big) \qquad \textit{if} \quad v_k < -\epsilon_k. \tag{2.27}$$

Proof. (i) We have $\langle p^k, u^{k+1} - \hat{u}^k \rangle = -t_k |p^k|^2$ by (2.18), whereas by (2.20),

$$\check{e}_k(u^{k+1}) = \bar{e}_C^k(u^{k+1}) = \bar{e}_C^k(\hat{u}^k) + \langle p^k, u^{k+1} - \hat{u}^k \rangle;$$

so $v_k := [h_{\hat{u}}^k]_+ - \check{e}_k(u^{k+1}) = \epsilon_k + t_k |p^k|^2$ by (2.18). Note that $v_k \geq \epsilon_k$.

(ii) This follows from (2.23) and the first part of (2.18).

(iii) By the definitions of $\bar{e}_C^k$ and ϵ_k in (2.17)–(2.18), we may express $-\epsilon_k$ as follows:

$$-\epsilon_k = \nu_k [\bar{f}_k(\hat{u}^k) - \tau_k] + (1 - \nu_k) \bar{h}_k(\hat{u}^k) + \bar{\imath}_C^k(\hat{u}^k) - [h_{\hat{u}}^k]_+,$$

where $\nu_k \in [0,1]$ by (2.13), $\bar{f}_k(\hat{u}^k) \leq f(\hat{u}^k) \leq f_{\hat{u}}^k + \epsilon_f$, $\bar{h}_k(\hat{u}^k) \leq h(\hat{u}^k) \leq h_{\hat{u}}^k + \epsilon_h$, and $\bar{\imath}_C^k(\hat{u}^k) \leq i_C(\hat{u}^k) = 0$ by (2.14)–(2.16) and (2.7), and $\tau_k \geq f_{\hat{u}}^k$ by (2.8). Therefore, we have

$$-\epsilon_k \leq \nu_k \epsilon_f + (1 - \nu_k) h(\hat{u}^k) - (1 - \nu_k) \big[h_{\hat{u}}^k\big]_+ \leq \nu_k \epsilon_f + (1 - \nu_k) \epsilon_h \leq \epsilon_{\max}.$$

(iv) Since $V_k \leq \max\{|p^k|, \epsilon_k\}(1 + |\hat{u}^k|)$ by (2.19) and the Cauchy–Schwarz inequality, the bounds follow from the equivalences in statement (ii), using $v_k \geq \epsilon_k$ and (2.24). ∎

The bound (2.27) will imply that if $\tau_k > f_*$ (so that $E(\tau_k) < 0$ by Lemma 2.1(vi), and hence V_k cannot vanish in (2.22) as t_k increases), then both $v_k \geq -\epsilon_k$ and the bound (2.26) must hold for t_k large enough.

2.6. Linearization selection. For choosing the sets J_f^{k+1} and J_h^{k+1}, note that (2.4)–(2.5) and (2.11) yield the existence of multipliers α_j^k for the pieces f_j, $j \in J_f^k$, and β_j^k for the pieces h_j, $j \in J_h^k$, such that

$$(2.28a) \qquad (p_f^k, 1) = \sum_{j \in J_f^k} \alpha_j^k (\nabla f_j, 1)\ \alpha_j^k \geq 0,\ \alpha_j^k \big[\check{f}_k(u^{k+1}) - f_j(u^{k+1})\big] = 0,\ j \in J_f^k,$$

$$(2.28b) \qquad (p_h^k, 1) = \sum_{j \in J_h^k} \beta_j^k (\nabla h_j, 1)\ \beta_j^k \geq 0,\ \beta_j^k \big[\check{h}_k(u^{k+1}) - h_j(u^{k+1})\big] = 0,\ j \in J_h^k.$$

Denote the indices of linearizations f_j and h_j that are "strongly" active at u^{k+1} by

$$(2.29) \qquad \hat{J}_f^k := \big\{j \in J_f^k : \alpha_j^k \neq 0\big\} \quad \text{and} \quad \hat{J}_h^k := \big\{j \in J_h^k : \beta_j^k \neq 0\big\}.$$

These linearizations embody all the information contained in the aggregates $\bar{f}_k$ and $\bar{h}_k$ (which are actually their convex combinations; cf. (2.14)–(2.15) and (2.28)). To save storage and work per iteration, we may drop the remaining linearizations. (Alternative strategies based on aggregation instead of selection are discussed in section 4.2.)

2.7. The method. We now have the necessary ingredients to state our method in detail.

ALGORITHM 2.6.

Step 0 (initialization). Select $u^1 \in C$, a *descent parameter* $\kappa \in (0,1)$, an *infeasibility contraction bound* $\kappa_h \in (0,1]$, a *stepsize bound* $t_{\min} > 0$, a stepsize $t_1 \geq t_{\min}$, and a penalty coefficient $c_1 \geq 0$. Set $\hat{u}^1 := u^1$, $f_{\hat{u}}^1 := f_u^1 := f_{u^1}$, $g_f^1 := g_f^{u^1}$, $h_{\hat{u}}^1 := h_u^1 := h_{u^1}$, $g_h^1 := g_h^{u^1}$ (cf. (2.4)), $J_f^1 := J_h^1 := \{1\}$, $i_t^1 := 0$, $k := k(0) := 1$, and $l := 0$ ($k(l) - 1$ will denote the iteration of the lth descent step).

Step 1 (trial point finding). For $\check{e}_k$ given by (2.8), find u^{k+1} (cf. (2.9)) and multipliers α_j^k, β_j^k such that (2.28) holds. Set v_k by (2.10), $p^k := (\hat{u}^k - u^{k+1})/t_k$, and $\epsilon_k := v_k - t_k|p^k|^2$.

Step 2 (stopping criterion). If $V_k = 0$ (cf. (2.19)) and $h_{\hat{u}}^k \le 0$, stop ($f_{\hat{u}}^k \le f_*$).

Step 3 (phase 1 stepsize correction). If $h_{\hat{u}}^k \le 0$ or $\epsilon_{\max} = 0$ or $v_k \ge \kappa_h h_{\hat{u}}^k$, go to Step 4. Set $t_k := 10t_k$, $i_t^k := k$. If $c_k > 0$, set $c_k := 2c_k$; otherwise, pick $c_k > 0$. Go back to Step 1.

Step 4 (stepsize correction). If $v_k \ge -\epsilon_k$, go to Step 5. Set $t_k := 10t_k$, $i_t^k := k$. If $h_{\hat{u}}^k > 0$, set $c_k := 2c_k$ if $c_k > 0$; otherwise, pick $c_k > 0$. Go back to Step 1.

Step 5 (descent test). Evaluate f_{k+1} and h_{k+1} (cf. (2.4)). If the *descent test* holds,

$$\max\{f_u^{k+1} - \tau_k, h_u^{k+1}\} \le [h_{\hat{u}}^k]_+ - \kappa v_k, \tag{2.30}$$

set $\hat{u}^{k+1} := u^{k+1}$, $f_{\hat{u}}^{k+1} := f_u^{k+1}$, $h_{\hat{u}}^{k+1} := h_u^{k+1}$, $i_t^{k+1} := 0$, and $k(l+1) := k+1$ and increase l by 1 (*descent step*); else set $\hat{u}^{k+1} := \hat{u}^k$, $f_{\hat{u}}^{k+1} := f_{\hat{u}}^k$, $h_{\hat{u}}^{k+1} := h_{\hat{u}}^k$, and $i_t^{k+1} := i_t^k$ (*null step*).

Step 6 (bundle selection). For the active sets $\hat{J}_f^k$ and $\hat{J}_h^k$ given by (2.29), choose

$$J_f^{k+1} \supset \hat{J}_f^k \cup \{k+1\} \quad \text{and} \quad J_h^{k+1} \supset \hat{J}_h^k \cup \{k+1\}. \tag{2.31}$$

Step 7 (stepsize updating). If $k(l) = k+1$ (i.e., after a descent step), select $t_{k+1} \ge t_k$ and $c_{k+1} \ge 0$; otherwise, set $c_{k+1} := c_k$ and either set $t_{k+1} := t_k$, or choose $t_{k+1} \in [t_{\min}, t_k]$ if $i_t^{k+1} = 0$.

Step 8 (loop). Increase k by 1 and go to Step 1.

Several comments on the method are in order.

Remark 2.7. (i) When the set C is polyhedral, Step 1 may use the QP method of [Kiw94], which can efficiently solve sequences of related subproblems (2.9).

(ii) Step 2 may also use the test $\inf \check{e}_C^k \ge 0$ and $h_{\hat{u}}^k \le 0$ (see Lemma 3.1(i) below).

(iii) Step 3 is needed in phase 1 (for $h_{\hat{u}}^k > 0$) when inaccuracies occur ($\epsilon_{\max} > 0$); it increases t_k and τ_k (via c_k) to obtain $v_k \ge \kappa_h h_{\hat{u}}^k$, so that eventually a descent step (cf. (2.30)) will reduce the constraint violation significantly: $h_{\hat{u}}^{k+1} \le (1 - \kappa\kappa_h) h_{\hat{u}}^k$.

(iv) In the case of exact evaluations ($\epsilon_{\max} = 0$), Step 4 is redundant, since $v_k \ge \epsilon_k \ge 0$ (cf. (2.23)–(2.24)). When inexactness is discovered via $v_k < -\epsilon_k$, t_k is increased to produce descent or confirm that $\hat{u}^k$ is almost optimal. Namely, when $\hat{u}^k$ is bounded in (2.27), increasing t_k drives V_k to 0, so that $f_{\hat{u}}^k \le \tau_k \le f_*$ asymptotically. Whenever t_k is increased at Steps 3 or 4, the *stepsize indicator* $i_t^k \ne 0$ prevents Step 7 from decreasing t_k after null steps until the next descent step occurs (cf. Step 5). Otherwise, decreasing t_k at Step 7 aims at collecting more local information about f and h at null steps.

(v) When $\epsilon_{\max} := \max\{\epsilon_f, \epsilon_h\} = 0$, our method employs the exact function values

$$f_{\hat{u}}^k = f(\hat{u}^k), \quad h_{\hat{u}}^k = h(\hat{u}^k), \quad \tau_k = \pi(\hat{u}^k; c_k) \ge f(\hat{u}^k), \quad \text{and} \quad [h_{\hat{u}}^k]_+ = e(\hat{u}^k; \tau_k) \tag{2.32}$$

(cf. (2.7), (2.1), (2.8), and Lemma 2.3), and the aggregate inequality (2.21) means that

$$p^k \in \partial_{\epsilon_k} e_C(\hat{u}^k; \tau_k) \quad \text{with} \quad \epsilon_k \ge 0. \tag{2.33}$$

Thus, if $V_k = 0$ in (2.19), then $|p^k| = \epsilon_k = 0$ implies that $0 \in \partial e_C(\hat{u}^k; \tau_k)$ and hence that $\hat{u}^k \in U_*$ by Lemma 2.2; in particular, in this case we have $h_{\hat{u}}^k = h(\hat{u}^k) \le 0$.

(vi) At Step 5, we have $v_k > 0$ (using (2.26) and $V_k > 0$ at Step 2 if $h^k_{\hat{u}} \le 0$; otherwise $v_k \ge \kappa_h h^k_{\hat{u}} > 0$ by Step 3 if $\epsilon_{\max} > 0$, $V_k > 0$ by item (v) if $\epsilon_{\max} = 0$). When a descent step occurs, the descent test (2.30) with the target τ_k given by (2.8) implies that

$$h^{k+1}_{\hat{u}} \le h^k_{\hat{u}} - \kappa v_k \qquad \text{if } h^k_{\hat{u}} > 0, \tag{2.34a}$$

$$f^{k+1}_{\hat{u}} \le f^k_{\hat{u}} - \kappa v_k \quad \text{and} \quad h^{k+1}_{\hat{u}} \le 0 \quad \text{if } h^k_{\hat{u}} \le 0. \tag{2.34b}$$

Thus at phase 1 (i.e., when $h^k_{\hat{u}} > 0$), we have reduction in the constraint violation, whereas at phase 2 the objective value is decreased while preserving (approximate) feasibility. In the exact case (cf. (2.32)), the descent test (2.30) becomes

$$\max\left\{f(u^{k+1}) - f(\hat{u}^k) - c_k h(\hat{u}^k)_+, h(u^{k+1})\right\} \le h(\hat{u}^k)_+ - \kappa v_k,$$

coinciding with the tests used in [Kiw85, section 5.7] and [KRSS07, SaS05] with $c_k \equiv 0$.

(vii) An active-set method for solving (2.9) (cf. [Kiw94]) will produce $|\hat{J}^k_f| + |\hat{J}^k_h| \le m + 1$ (cf. (2.29)). Hence Step 6 can keep $|J^{k+1}_f| + |J^{k+1}_h| \le \bar{m}$ for any given bound $\bar{m} \ge m + 3$.

(viii) Step 7 may use the techniques of [Kiw90, LeS97] for updating t_k (or the proximity weight $1/t_k$) with obvious modifications. For updates of c_k, see section 4.4.

3. Convergence. Our analysis splits into several cases.

3.1. The case of an infinite cycle due to oracle errors. We first show that, in phase 2, the loop between Steps 1 and 4 is infinite iff $0 \le \inf \check{e}^k_C < \check{e}_k(\hat{u}^k)$, in which case $\hat{u}^k$ is *approximately optimal*: $f(\hat{u}^k) \le f_* + \epsilon_f$ and $h(\hat{u}^k) \le \epsilon_h$.

LEMMA 3.1. *Assuming that $h^k_{\hat{u}} \le 0$, recall that $\check{E}_k := \inf \check{e}^k_C$ with $\check{e}^k_C := \check{e}_k + i_C$. Then we have the following statements*:

(i) *If $\check{E}_k \ge 0$, then $f(\hat{u}^k) - \epsilon_f \le f^k_{\hat{u}} \le f_*$ and $h(\hat{u}^k) \le \epsilon_h$.*

(ii) *Step 2 terminates, i.e., $V_k := \max\{|p^k|, \epsilon_k + \langle p^k, \hat{u}^k\rangle\} = 0$, iff $0 \le \check{E}_k = \check{e}_k(\hat{u}^k)$.*

(iii) *If the loop between Steps 1 and 4 is infinite, then $\check{E}_k \ge 0$ and $V_k \to 0$.*

(iv) *If $\check{E}_k \ge 0$ at Step 1 and Step 2 does not terminate (i.e., $\check{E}_k < \check{e}_k(\hat{u}^k)$; cf. (ii)), then an infinite loop between Steps 4 and 1 occurs.*

Proof. (i) We have $E(\tau_k) \ge \check{E}_k$ and $\tau_k = f^k_{\hat{u}}$ (cf. (2.2), (2.8), (2.14)–(2.15)); so $f^k_{\hat{u}} \le f_*$ by Lemma 2.1(vi), whereas $f(\hat{u}^k) \le f^k_{\hat{u}} + \epsilon_f$ and $h(\hat{u}^k) \le h^k_{\hat{u}} + \epsilon_h$ by (2.7).

(ii) "⇒": Since $|p^k| = 0 \ge \epsilon_k$, (2.18) and (2.21) yield $u^{k+1} = \hat{u}^k$, $\bar{e}^k_C(\hat{u}^k) \le \check{e}^k_C(\cdot)$ and $0 \le \bar{e}^k_C(\hat{u}^k)$, whereas by (2.20), $\bar{e}^k_C(\hat{u}^k) = \check{e}_k(u^{k+1}) = \check{e}_k(\hat{u}^k)$. "⇐": Since $\check{e}^k_C(\hat{u}^k) = \min \check{e}^k_C$, using $\phi_k(\hat{u}^k) = \min \check{e}^k_C \le \phi_k(u^{k+1}) \le \phi_k(\hat{u}^k)$ in (2.9) gives $u^{k+1} = \hat{u}^k$; thus $\bar{e}^k_C(\hat{u}^k) = \check{e}^k_C(\hat{u}^k)$ by (2.20), and (2.18) yields $p^k = 0$ and $\epsilon_k = -\check{e}^k_C(\hat{u}^k) \le 0$.

(iii) At Step 4 during the loop the facts that $V_k < (2\epsilon_{\max}/t_k)^{1/2}(1 + |\hat{u}^k|)$ (cf. (2.27)) and $t_k \uparrow \infty$ as the loop continues give $V_k \to 0$; so $\check{e}^k_C(\cdot) \ge 0$ by (2.22).

(iv) We have $\check{e}_k(u^{k+1}) \ge \inf \check{e}^k_C \ge 0$. Thus $v_k = -\check{e}_k(u^{k+1}) \le 0$ (cf. (2.10)) and $v_k = t_k|p^k|^2 + \epsilon_k$ (cf. (2.23)) yield $\epsilon_k \le -t_k|p^k|^2$ at Step 4 with $p^k \ne 0$ (since $\max\{|p^k|, \epsilon_k + \langle p^k, \hat{u}^k\rangle\} =: V_k > 0$ at Step 2). Hence $\epsilon_k < -\frac{t_k}{2}|p^k|^2$; so $v_k < -\epsilon_k$ and Step 4 loops back to Step 1, after which Step 2 cannot terminate due to (ii). ∎

In view of Lemma 3.1, from now on we assume (unless stated otherwise) that the algorithm neither terminates nor cycles infinitely between Steps 1 and 4 at phase 2 (otherwise $\hat{u}^k$ is approximately optimal). For phase 1, our analysis will imply that any loop between Steps 1 and 3 or 4 is finite. We shall show that the algorithm generates points that are approximately optimal asymptotically by establishing upper bounds on the values $f^k_{\hat{u}}$ and $h^k_{\hat{u}}$.

3.2. Bounding the objective values. We first bound $f_{\hat{u}}^k$ via V_k.

LEMMA 3.2. *Let $K \subset \mathbb{N}$ satisfy $V_k \xrightarrow{K} 0$. Then $\overline{\lim}_{k\in K} f_{\hat{u}}^k \le \overline{\lim}_{k\in K} \tau_k \le f_*$.*

Proof. Pick $K' \subset K$ such that $\tau_k \xrightarrow{K'} \bar{\tau} := \overline{\lim}_{k\in K} \tau_k$. Since $f_{\hat{u}}^k \le \tau_k$ by (2.8), we need only show that $\bar{\tau} \le f_*$ when $\bar{\tau} > -\infty$. Note that $\bar{\tau} < \infty$, since otherwise for $\tau_k \ge f(\mathring{u}) - h(\mathring{u})$, the fact that $e(\mathring{u};\tau_k) = h(\mathring{u}) < 0$ (cf. (2.2), (1.2)) and the bound (2.22) would yield the following contradiction:

$$0 > h(\mathring{u}) = e_C(\mathring{u};\tau_k) \ge -V_k(1+|\mathring{u}|) \xrightarrow{K'} 0.$$

Thus $\bar{\tau}$ is finite. Since $e_C(u;\cdot)$ is continuous, letting $k \xrightarrow{K'} \infty$ in (2.22) gives $e_C(\cdot;\bar{\tau}) \ge 0$. Therefore, we have $E(\bar{\tau}) \ge 0$, and hence $\bar{\tau} \le f_*$ by Lemma 2.1(vi). ∎

The upper bound of Lemma 3.2 is complemented below with a lower bound (which is highly useful for the "dual" applications in sections 4.3 and 5).

LEMMA 3.3. *If $\overline{\lim}_k h_{\hat{u}}^k \le 0$, then for the minimal multiplier $\bar{\mu} := \inf_{\mu\in M} \mu$ of problem (1.1) (cf. section 2.1), we have*

$$\underline{\lim}_k f_{\hat{u}}^k + \epsilon_f \ge \underline{\lim}_k f(\hat{u}^k) \ge f_* - \bar{\mu}\epsilon_h \quad \text{and} \quad \overline{\lim}_k h(\hat{u}^k) \le \epsilon_h. \tag{3.1}$$

Proof. For all k, $\hat{u}^k \in C$ and (cf. section 2.1) $L(\hat{u}^k;\bar{\mu}) := f(\hat{u}^k) + \bar{\mu}h(\hat{u}^k) \ge f_*$, with $0 \le \bar{\mu} < \infty$ if $f_* > -\infty$, $\bar{\mu} = \infty$ otherwise. Moreover, $f(\hat{u}^k) \le f_{\hat{u}}^k + \epsilon_f$, and $h(\hat{u}^k) \le h_{\hat{u}}^k + \epsilon_h$ by (2.7). The conclusion follows. ∎

3.3. The case of finitely many descent steps. We now consider the case where only finitely many descent steps occur. After the last descent step, only null steps occur and $\{t_k\}$ becomes eventually monotone, since once Steps 3 or 4 increase t_k, Step 7 cannot decrease t_k; thus the limit $t_\infty := \lim_k t_k$ exists. After showing that $t_\infty = \infty$ may occur only at phase 2 in Lemma 3.4, we deal with the cases of $t_\infty = \infty$ in Lemma 3.6 and $t_\infty < \infty$ in Lemma 3.7.

LEMMA 3.4. *Suppose there exists $\bar{k}$ such that $h_{\hat{u}}^{\bar{k}} > 0$ and only null steps occur for all $k \ge \bar{k}$. Then Steps 3 and 4 can increase t_k only a finite number of times.*

Proof. For contradiction, suppose that $t_k \to \infty$. Since $\tau_k \to \infty$ (because $c_k \to \infty$; cf. Steps 3 and 4 and (2.8)), we may assume that $\tau_k \ge \mathring{\tau} := f(\mathring{u}) - h(\mathring{u})$ for the Slater point $\mathring{u}$ of (1.2) and for all $k \ge \bar{k}$; then, using the minorants $\check{f}_k \le f$ and $\check{h}_k \le h$ (cf. (2.4)) in the definitions (2.8) and (2.2) yields

$$\check{e}_k(\mathring{u}) \le \max\{\check{f}_k(\mathring{u}) - \mathring{\tau}, \check{h}_k(\mathring{u})\} \le e(\mathring{u};\mathring{\tau}) = h(\mathring{u}) < 0 \quad \text{with} \quad \mathring{u} \in C. \tag{3.2}$$

At Step 1, (2.9) gives the proximal projection property for the level set of $\check{e}_C^k := \check{e}_k + i_C$:

$$u^{k+1} = \arg\min\{\tfrac{1}{2}|u - \hat{u}^k|^2 : \check{e}_C^k(u) \le \check{e}_C^k(u^{k+1})\}, \tag{3.3}$$

whereas before Step 3 increases t_k, $v_k < \kappa_h h_{\hat{u}}^k$ yields $\check{e}_k(u^{k+1}) > (1-\kappa_h)h_{\hat{u}}^k \ge 0$ by (2.10); so for $k \ge \bar{k}$, (3.2) and (3.3) with $\hat{u}^k = \hat{u}^{\bar{k}}$ give $|u^{k+1} - \hat{u}^k| \le r := |\mathring{u} - \hat{u}^{\bar{k}}|$, and hence $|p^k| \le r/t_k$ by (2.18). Therefore, if Step 3 increases t_k at infinitely many iterations, indexed by K, say, then $t_k \to \infty$ yields $p^k \xrightarrow{K} 0$; thus, from (2.21), (2.20), the fact that $|u^{k+1} - \hat{u}^{\bar{k}}| \le r$, and the Cauchy–Schwarz inequality, we get

$$0 > h(\mathring{u}) \ge \check{e}_C^k(\mathring{u}) \ge \bar{e}_C^k(\mathring{u}) = \check{e}_k(u^{k+1}) + \langle p^k, \mathring{u} - u^{k+1}\rangle \ge \langle p^k, \mathring{u} - u^{k+1}\rangle \xrightarrow{K} 0,$$

a contradiction. Similarly, if Step 4 is entered with $v_k < -\epsilon_k$ for infinitely many iterations indexed by K, say, then $t_k \to \infty$ and (2.27) give $V_k \xrightarrow{K} 0$, and we obtain

$$0 > h(\mathring{u}) \ge \check{e}_C^k(\mathring{u}) \ge -V_k(1+|\mathring{u}|) \xrightarrow{K} 0$$

from (3.2) and (2.22), another contradiction. The conclusion follows. ∎

Remark 3.5. To illustrate the need for increasing c_k at Steps 3 and 4, suppose momentarily that $c_k \equiv 0$ for all k. Consider the following example. Let $m = 1$, $f(u) := u$, $h(u) := 1 - u$, $C := R$. Suppose that $u^1 := 0$, $f_1 := f$, $h_1 := h - 0.5$; so that $h^1_{\hat{u}} = 0.5$ for $\epsilon_h = 0.5$. For $k = 1$, $v_k \le 1/4$; so if $\kappa_h \in (1/2, 1)$, then a loop between Steps 3 and 1 occurs. Next, for $\kappa_h \in (0, 1/2]$, suppose $f_{k+1} = f$ and $h_{k+1} = h$ at Step 5; then a null step occurs, and at Step 1 for $k = 2$, $\check{e}_k = \max\{f, h\}$ is exact, $\min \check{e}_k = 1/2 = h^k_{\hat{u}}$, and $v_k \le 0$, so that a loop between Steps 3 and 1 occurs. Even if Step 3 were omitted, a loop between Steps 4 and 1 would occur.

The case where the stepsize t_k keeps growing at a fixed prox center is quite simple.

LEMMA 3.6. *Suppose there exists $\bar{k}$ such that only null steps occur for all $k \ge \bar{k}$, and $t_\infty := \lim_k t_k = \infty$. Let $K := \{k \ge \bar{k} : t_{k+1} > t_k\}$. Then $V_k \xrightarrow{K} 0$, and $h^{\bar{k}}_{\hat{u}} \le 0$.*

Proof. We have $h^{\bar{k}}_{\hat{u}} \le 0$ (otherwise Lemma 3.4 would imply $t_\infty < \infty$, a contradiction). For $k \in K$, before t_k is increased at Step 4 on the last loop to Step 1, we have $V_k < (2\epsilon_{\max}/t_k)^{1/2}(1 + |\hat{u}^{\bar{k}}|)$ by (2.27); so $t_k \to \infty$ gives $V_k \xrightarrow{K} 0$. ∎

The case where the stepsize t_k does not grow at a fixed prox center is analyzed as in [Kiw06a]. After showing that the optimal value $\phi_k(u^{k+1})$ of subproblem (2.9) is nondecreasing and bounded above, u^{k+1} is bounded, and $u^{k+2} - u^{k+1} \to 0$, we invoke the descent test (2.30) to get $v_k \to 0$; the rest follows from the bounds (2.25)–(2.26).

LEMMA 3.7. *Suppose that there exists $\bar{k}$ such that for all $k \ge \bar{k}$, only null steps occur, and Steps 3 and 4 do not increase t_k. Then $V_k \to 0$, and $h^{\bar{k}}_{\hat{u}} \le 0$.*

Proof. Fix $k \ge \bar{k}$. We show that the aggregate $\bar{e}^k_C$ minorizes the next model $\check{e}^{k+1}_C$:

$$\bar{e}^k_C(\cdot) \le \check{e}^{k+1}_C(\cdot) := \check{e}_{k+1}(\cdot) + i_C(\cdot). \tag{3.4}$$

Consider the selected model $\hat{f}_k := \max_{j \in \hat{J}^k_f} f_j$ of $\check{f}_k := \max_{j \in J^k_f} f_j$; then $\hat{f}_k \le \check{f}_k$. Using (2.29) in the expression (2.28a) of p^k_f gives $\hat{f}_k(u^{k+1}) = \check{f}_k(u^{k+1})$ and $p^k_f \in \partial \hat{f}_k(u^{k+1})$ (cf. [HUL93, Ex. VI.3.4]). Thus $\bar{f}_k \le \hat{f}_k$ by (2.14); so the choice of $\hat{J}^k_f \subset J^{k+1}_f$ implies that $\bar{f}_k \le \hat{f}_k \le \check{f}_{k+1}$. Similarly, for $\hat{h}_k := \max_{j \in \hat{J}^k_h} h_j$, (2.28b) yields $\bar{h}_k \le \hat{h}_k \le \check{h}_{k+1}$. Then using the definition (2.17) of $\bar{e}^k_C$ with $\nu_k \in [0, 1]$ (cf. (2.13)), the minorization $i^k_C \le i_C$ of (2.16), and the fact that $\tau_{k+1} = \tau_k$ (by (2.8) and Steps 3 and 4) gives the required bound

$$\bar{e}^k_C \le \nu_k[\check{f}_{k+1} - \tau_k] + (1 - \nu_k)\check{h}_{k+1} + i_C \le \max\{\check{f}_{k+1} - \tau_{k+1}, \check{h}_{k+1}\} + i_C = \check{e}^{k+1}_C.$$

(Note that this bound needs only the minorizations $\bar{f}_k \le \check{f}_{k+1} + i_C$ and $\bar{h}_k \le \check{h}_{k+1} + i_C$; this will be important when selection is replaced by aggregation in section 4.2.)

Next, consider the following partial linearization of the objective ϕ_k of (2.9):

$$\bar{\phi}_k(\cdot) := \bar{e}^k_C(\cdot) + \tfrac{1}{2t_k}|\cdot - \hat{u}^k|^2. \tag{3.5}$$

We have $\bar{e}^k_C(u^{k+1}) = \check{e}_k(u^{k+1})$ by (2.20) and $\nabla \bar{\phi}_k(u^{k+1}) = 0$ from $\nabla \bar{e}^k_C = p^k = (\hat{u}^k - u^{k+1})/t_k$ (cf. (2.20), (2.18)); hence $\bar{\phi}_k(u^{k+1}) = \phi_k(u^{k+1})$ by (2.9), and by Taylor's expansion

$$\bar{\phi}_k(\cdot) = \phi_k(u^{k+1}) + \tfrac{1}{2t_k}|\cdot - u^{k+1}|^2. \tag{3.6}$$

To bound $\bar{\phi}_k(\hat{u}^k)$ from above, notice that (3.5), (2.18), and (2.24) imply that

$$\bar{\phi}_k(\hat{u}^k) = \bar{e}^k_C(\hat{u}^k) = [h^k_{\hat{u}}]_+ - \epsilon_k \le [h^k_{\hat{u}}]_+ + \epsilon_{\max}.$$

Then by (3.6),

$$\phi_k(u^{k+1}) + \tfrac{1}{2t_k}|u^{k+1} - \hat{u}^k|^2 = \bar{\phi}_k(\hat{u}^k) \le \big[h^{\bar{k}}_{\hat{u}}\big]_+ + \epsilon_{\max}. \tag{3.7}$$

Now using the facts that $\hat{u}^{k+1} = \hat{u}^k$ and $t_{k+1} \le t_k$ and the model minorization property (3.4) in the definitions (3.5) of $\bar{\phi}_k$ and (2.9) of ϕ_{k+1} gives $\bar{\phi}_k \le \phi_{k+1}$. Hence by (3.6),

$$\phi_k(u^{k+1}) + \tfrac{1}{2t_k}|u^{k+2} - u^{k+1}|^2 = \bar{\phi}_k(u^{k+2}) \le \phi_{k+1}(u^{k+2}). \tag{3.8}$$

Thus the nondecreasing sequence $\{\phi_k(u^{k+1})\}_{k\ge\bar{k}}$, being bounded above by (3.7) with $\hat{u}^k = \hat{u}^{\bar{k}}$ for $k \ge \bar{k}$, must have a limit, say $\phi_\infty \le [h^{\bar{k}}_{\hat{u}}]_+ + \epsilon_{\max}$. Moreover, since the stepsizes satisfy $t_k \le t_{\bar{k}}$ for $k \ge \bar{k}$, we deduce from the bounds (3.7)–(3.8) that

$$\phi_k(u^{k+1}) \uparrow \phi_\infty, \quad u^{k+2} - u^{k+1} \to 0, \tag{3.9}$$

and the sequence $\{u^{k+1}\}$ is bounded. Then the sequence $\{g_f^{k+1}\}$ is bounded as well, since $g_f^k \in \partial_{\epsilon_f} f(u^k)$ by (2.4), whereas the mapping $\partial_{\epsilon_f} f$ is locally bounded [HUL93, section XI.4.1]; similarly, the sequence $\{g_h^{k+1}\}$ is bounded, since $g_h^k \in \partial_{\epsilon_h} h(u^k)$ by (2.4).

For $v_k := [h^k_{\hat{u}}]_+ - \check{e}_k(u^{k+1})$ and the following linearization of $e(\cdot;\tau_k)$ at u^{k+1},

$$e_{k+1}(\cdot) := \begin{cases} f_{k+1}(\cdot) - \tau_k & \text{if } f_u^{k+1} - \tau_k \ge h_u^{k+1}, \\ h_{k+1}(\cdot) & \text{otherwise,} \end{cases} \tag{3.10}$$

the descent test (2.30) reads $e_{k+1}(u^{k+1}) \le [h^k_{\hat{u}}]_+ - \kappa v_k$ or equivalently

$$\tilde{\epsilon}_k := e_{k+1}(u^{k+1}) - \check{e}_k(u^{k+1}) \le (1-\kappa)v_k. \tag{3.11}$$

We now show that this approximation error $\tilde{\epsilon}_k \to 0$. First, note that the linearization gradients $g_e^{k+1} := \nabla e_{k+1}$ are bounded, since $|g_e^{k+1}| \le \max\{|g_f^{k+1}|, |g_h^{k+1}|\}$ by (2.4). Further, the minorizations $f_{k+1} \le \check{f}_{k+1}$ and $h_{k+1} \le \check{h}_{k+1}$ due to $k+1 \in J_f^{k+1} \cap J_h^{k+1}$ (cf. (2.5)) yield $e_{k+1} \le \check{e}_{k+1}$ by (2.8), since $\tau_{k+1} = \tau_k$. Using the linearity of e_{k+1}, the bound $e_{k+1} \le \check{e}_{k+1}$, the Cauchy–Schwarz inequality, and (2.9) with $\hat{u}^k = \hat{u}^{\bar{k}}$ for $k \ge \bar{k}$, we estimate

$$\begin{aligned} \tilde{\epsilon}_k &:= e_{k+1}(u^{k+1}) - \check{e}_k(u^{k+1}) \\ &= e_{k+1}(u^{k+2}) - \check{e}_k(u^{k+1}) + \langle g_e^{k+1}, u^{k+1} - u^{k+2}\rangle \\ &\le \check{e}_{k+1}(u^{k+2}) - \check{e}_k(u^{k+1}) + |g_e^{k+1}||u^{k+1} - u^{k+2}| \\ &= \phi_{k+1}(u^{k+2}) - \phi_k(u^{k+1}) + \Delta_k + |g_e^{k+1}||u^{k+1} - u^{k+2}|, \end{aligned} \tag{3.12}$$

where $\Delta_k := |u^{k+1} - \hat{u}^{\bar{k}}|^2/2t_k - |u^{k+2} - \hat{u}^{\bar{k}}|^2/2t_{k+1}$. We have $\Delta_k \to 0$, since $t_{\min} \le t_{k+1} \le t_k$ (cf. Step 7), $|u^{k+1} - \hat{u}^{\bar{k}}|^2$ is bounded, $u^{k+2} - u^{k+1} \to 0$ by (3.9), and

$$|u^{k+2} - \hat{u}^{\bar{k}}|^2 = |u^{k+1} - \hat{u}^{\bar{k}}|^2 + 2\langle u^{k+2} - u^{k+1}, u^{k+1} - \hat{u}^{\bar{k}}\rangle + |u^{k+2} - u^{k+1}|^2.$$

Hence, using (3.9) and the boundedness of $\{g_e^{k+1}\}$ in (3.12) yields $\overline{\lim}_k \tilde{\epsilon}_k \le 0$. On the other hand, for $k \ge \bar{k}$, the descent test written as (3.11) fails: $(1-\kappa)v_k < \tilde{\epsilon}_k$, where $\kappa < 1$ and $v_k > 0$; it follows that $\tilde{\epsilon}_k \to 0$ and $v_k \to 0$.

Since $v_k \to 0$, $t_k \geq t_{\min}$, and $\hat{u}^k = \hat{u}^{\bar{k}}$ for $k \geq \bar{k}$, we have $V_k \to 0$ by (2.26), $\epsilon_k \to 0$, and $|p^k| \to 0$ by (2.25). It remains to prove that $h_{\hat{u}}^{\bar{k}} \leq 0$. If $\epsilon_{\max} > 0$, but $h_{\hat{u}}^{\bar{k}} > 0$, then the facts that $v_k \to 0$ with $v_k \geq \kappa_h h_{\hat{u}}^k$ (cf. Step 3), $\kappa_h > 0$, and $h_{\hat{u}}^k = h_{\hat{u}}^{\bar{k}}$ for $k \geq \bar{k}$ give in the limit $h_{\hat{u}}^{\bar{k}} \leq 0$, a contradiction. Finally, for $\epsilon_{\max} = 0$, recalling Remark 2.7(v) and using $\epsilon_k, |p^k| \to 0$ in (2.21) yields $e_C(\hat{u}^{\bar{k}}; \tau_{\bar{k}}) \leq e_C(\cdot; \tau_{\bar{k}})$. In other words, we have $0 \in \partial e_C(\hat{u}^{\bar{k}}; \tau_{\bar{k}})$; so $\hat{u}^{\bar{k}} \in U_*$ by Lemma 2.2, and thus $h_{\hat{u}}^{\bar{k}} = h(\hat{u}^{\bar{k}}) \leq 0$. ∎

We may now finish the case of infinitely many consecutive null steps.

THEOREM 3.8. *Suppose there exists $\bar{k}$ such that only null steps occur for all $k \geq \bar{k}$. Let $K := \{k \geq \bar{k} : t_{k+1} > t_k\}$ if $t_k \to \infty$, $K := \{k : k \geq \bar{k}\}$ otherwise. Then $V_k \xrightarrow{K} 0$, $f_{\hat{u}}^{\bar{k}} \leq f_*$ and $h_{\hat{u}}^{\bar{k}} \leq 0$. Moreover, the bounds of* (3.1) *hold.*

Proof. Steps 3, 4, 5, and 7 ensure that $\{t_k\}$ is monotone for large k (see above Lemma 3.4). We have $V_k \xrightarrow{K} 0$ and $h_{\hat{u}}^{\bar{k}} \leq 0$ from either Lemma 3.6 if $t_\infty = \infty$ or Lemma 3.7 if $t_\infty < \infty$. Then $f_{\hat{u}}^{\bar{k}} \leq f_*$ by Lemma 3.2 (since $\tau_k = f_{\hat{u}}^k = f_{\hat{u}}^{\bar{k}}$ for $k \geq \bar{k}$). The final assertion stems from Lemma 3.3. ∎

It may be interesting to observe that $u^k \to \hat{u}^{\bar{k}}$ if $t_\infty < \infty$ (since $|u^{k+1} - \hat{u}^k| = t_k|p^k|$ by (2.18), and $p^k \to 0$ in the proof of Lemma 3.7). In contrast, we may have $t_\infty = \infty$ and $|u^k| \to \infty$ (consider $m = 1$, $f(u) := e^u$, $h(u) \equiv -1$, $C := \mathbb{R}$, $u^1 := 0$, $f_u^1 := -1$, $g_f^1 = 1$, and exact evaluations for $k \geq 2$).

3.4. The case of infinitely many descent steps. We first analyze the case of infinitely many descent steps in phase 2.

THEOREM 3.9. *Suppose infinitely many descent steps occur, and $h_{\hat{u}}^{\bar{k}} \leq 0$ for some $\bar{k}$. Let $f_{\hat{u}}^\infty := \lim_k f_{\hat{u}}^k$ and $K := \{k \geq \bar{k} : f_{\hat{u}}^{k+1} < f_{\hat{u}}^k\}$. Then either $f_{\hat{u}}^\infty = f_* = -\infty$, or $-\infty < f_{\hat{u}}^\infty \leq f_*$ and $\underline{\lim}_{k \in K} V_k = 0$. Moreover, the bounds of* (3.1) *hold. In particular, if $\{\hat{u}^k\}$ is bounded, then $f_{\hat{u}}^\infty > -\infty$ and $V_k \xrightarrow{K} 0$.*

Proof. For $k \geq \bar{k}$, we have $h_{\hat{u}}^k \leq 0$, $\tau_k = f_{\hat{u}}^k$ (cf. (2.8)), and $f_{\hat{u}}^{k+1} \leq f_{\hat{u}}^k$, since by (2.34b), a descent step yields $h_{\hat{u}}^{k+1} \leq 0$ and $f_{\hat{u}}^{k+1} - f_{\hat{u}}^k \leq -\kappa v_k < 0$, so that $|K| = \infty$. First, suppose that $f_{\hat{u}}^\infty > -\infty$.

We have $0 < \kappa v_k \leq f_{\hat{u}}^k - f_{\hat{u}}^{k+1}$ if $k \in K$, $f_{\hat{u}}^{k+1} = f_{\hat{u}}^k$ otherwise; so $\sum_{k \in K} \kappa v_k \leq f_{\hat{u}}^{\bar{k}} - f_{\hat{u}}^\infty < \infty$ gives $v_k \xrightarrow{K} 0$ and hence $\epsilon_k, t_k|p^k|^2 \xrightarrow{K} 0$ by (2.25), as well as $|p^k| \xrightarrow{K} 0$, using $t_k \geq t_{\min}$. Now, for the descent iterations $k \in K$, we have $\hat{u}^{k+1} - \hat{u}^k = -t_k p^k$ by (2.18) and therefore

$$|\hat{u}^{k+1}|^2 - |\hat{u}^k|^2 = t_k\big\{t_k|p^k|^2 - 2\langle p^k, \hat{u}^k\rangle\big\}.$$

Sum up and use the facts that $\hat{u}^{k+1} = \hat{u}^k$ if $k \notin K$ and $\sum_{k \in K} t_k \geq \sum_{k \in K} t_{\min} = \infty$ to get

$$\overline{\lim_{k \in K}}\big\{t_k|p^k|^2 - 2\langle p^k, \hat{u}^k\rangle\big\} \geq 0$$

(since otherwise $|\hat{u}^k|^2 \to -\infty$, which is impossible). Combining this with $t_k|p^k|^2 \xrightarrow{K} 0$ gives $\underline{\lim}_{k \in K}\langle p^k, \hat{u}^k\rangle \leq 0$. Since also $\epsilon_k, |p^k| \xrightarrow{K} 0$, we have $\underline{\lim}_{k \in K} V_k = 0$ by (2.19).

Then using $\underline{\lim}_{k \in K} V_k = 0$ and $\tau_k \to f_{\hat{u}}^\infty$ in Lemma 3.2 shows that $f_{\hat{u}}^\infty \leq f_*$.

For the case of $f_{\hat{u}}^\infty = -\infty$ and the assertion on (3.1), invoke Lemma 3.3.

For the final assertion, if $\{\hat{u}^k\} \subset C$ is bounded, then $\inf_k f(\hat{u}^k) > -\infty$ (f is closed on C) implies that $f_{\hat{u}}^\infty > -\infty$ by (3.1); so we have $\epsilon_k, |p^k| \xrightarrow{K} 0$ as above. Hence the fact that $V_k \leq \max\{|p^k|, \epsilon_k\}(1 + |\hat{u}^k|)$ by Lemma 2.5(iv) gives $V_k \xrightarrow{K} 0$. ∎

We now deal with the case of infinitely many descent steps at phase 1 for $\epsilon_{\max} > 0$.

THEOREM 3.10. *Suppose infinitely many descent steps occur, $h^k_{\hat{u}} > 0$ for all k, and $\epsilon_{\max} > 0$. Let $K := \{k : h^{k+1}_{\hat{u}} < h^k_{\hat{u}}\}$. Then we have the following statements*:

(i) $h^k_{\hat{u}} \downarrow 0$ (*this relies upon the property that $v_k \geq \kappa_h h^k_{\hat{u}}$ at Step* 5).

(ii) $\underline{\lim}_{k\in K} V_k = 0$; *also* $\sum_{k\in K} v_k < \infty$, *and* $\lim_{k\in K} \max\{\epsilon_k, |p^k|\} = 0$.

(iii) *Let $K' \subset \mathbb{N}$ be such that $V_k \xrightarrow{K'} 0$. Then $\overline{\lim}_{k\in K'} f^k_{\hat{u}} \leq \overline{\lim}_{k\in K'} \tau_k \leq f_*$.*

(iv) *If $\{\hat{u}^k\}$ is bounded, then $\lim_{k\in K} V_k = 0$, and we may take $K' = K$ in* (iii).

(v) *The bounds of* (3.1) *hold, and $\underline{\lim}_k \tau_k \geq f_* - \epsilon_f - \bar{\mu}\epsilon_h$.*

(vi) *Assertions* (ii)–(iv) *above hold also if $\epsilon_{\max} = 0$.*

Proof. We have $h^{k+1}_{\hat{u}} - h^k_{\hat{u}} \leq -\kappa v_k < 0$ at descent steps by (2.34a); thus $|K| = \infty$.

(i) We have $0 < \kappa v_k \leq h^k_{\hat{u}} - h^{k+1}_{\hat{u}}$ if $k \in K$, $h^{k+1}_{\hat{u}} = h^k_{\hat{u}}$ otherwise; so $\sum_{k\in K} \kappa v_k \leq h^1_{\hat{u}}$ gives $\lim_{k\in K} v_k = 0$. Hence the fact that $v_k \geq \kappa_h h^k_{\hat{u}}$ (cf. Step 3) yields $h^k_{\hat{u}} \downarrow 0$.

(ii) Use $\sum_{k\in K} v_k < \infty$, and then $v_k \xrightarrow{K} 0$ (from the proof of (i)) as in the proof of Theorem 3.9 to get $\underline{\lim}_{k\in K} V_k = 0$, $\lim_{k\in K} \epsilon_k = 0$, and $\lim_{k\in K} |p^k| = 0$.

(iii) This follows from Lemma 3.2.

(iv) Invoke Lemma 2.5(iv) and the fact that $\lim_{k\in K} \max\{\epsilon_k, |p^k|\} = 0$ by (ii).

(v) This follows from (i), Lemma 3.3, and the fact that $\tau_k \geq f^k_{\hat{u}}$ for all k.

(vi) This statement is immediate from the preceding arguments and the rules of Step 3. ∎

It is instructive to examine the assumptions of the preceding results.

Remark 3.11. (i) Inspection of the preceding proofs reveals that Theorems 3.8–3.10 require only convexity and finiteness of f and h on C and *local boundedness* of the approximate subgradient mappings g^u_f of f and g^u_h of h on C. In particular, it suffices to assume that f and h are finite convex on a neighborhood of C.

(ii) Using the *evaluation errors* $\epsilon^k_f := f(u^k) - f^k_u$ and $\epsilon^k_h := h(u^k) - h^k_u$, our results are sharpened as follows; cf. [Kiw06b, section 4.2]. In general, $f(\hat{u}^k) = f^k_{\hat{u}} + \epsilon^{k(l)}_f$ and $h(\hat{u}^k) = h^k_{\hat{u}} + \epsilon^{k(l)}_h$, where $k(l) - 1$ denotes the iteration number of the lth descent step. Hence ϵ_f and ϵ_h in the bounds of (3.1) for Theorems 3.8–3.10 may be replaced by the *asymptotic errors* ϵ^∞_f and ϵ^∞_h, where ϵ^∞_f equals the final $\epsilon^{k(l)}_f$ if only finitely many descent steps occur, $\overline{\lim}_l \epsilon^{k(l)}_f$ otherwise, and ϵ^∞_h is defined analogously.

(iii) Concerning Theorem 3.10(iv), note that the sequence $\{\hat{u}^k\}$ is bounded if the feasible set U is bounded. Indeed, $h(\hat{u}^k) \leq h^k_{\hat{u}} + \epsilon_h$ (cf. (2.7)) with $h^k_{\hat{u}} \leq h^1_{\hat{u}}$ implies that $\{\hat{u}^k\}$ lies in the set $\{u \in C : h(u) \leq h^1_{\hat{u}} + \epsilon_h\}$, which is bounded, since such is U.

Finally, we analyze infinitely many descent steps in the exact case of $\epsilon_{\max} = 0$.

THEOREM 3.12. *Suppose that infinitely many descent steps occur and $\epsilon_{\max} = 0$. Let $K := \{k(l) - 1\}_{l=1}^\infty$ index the descent iterations* (*cf. Step* 5), *and let $\bar{k} := \inf\{k : h(\hat{u}^k) \leq 0\}$* (*so that phase* 2 *starts at iteration $k = \bar{k}$ iff $\bar{k} < \infty$*). *Then we have the following statements*:

(i) *If $\bar{k} < \infty$, then $f(\hat{u}^k) \to f_*$, $\tau_k \to f_*$, $h(\hat{u}^k)_+ \to 0$, and each cluster point of $\{\hat{u}^k\}$* (*if any*) *lies in the optimal set U_*; moreover, $\underline{\lim}_{k\in K} V_k = 0$ if $f_* > -\infty$.*

(ii) *If $\inf_k f(\hat{u}^k) > -\infty$ or $\bar{k} = \infty$, then $\sum_{k\in K} v_k < \infty$, $\epsilon_k \xrightarrow{K} 0$, and $p^k \xrightarrow{K} 0$.*

(iii) *If the sequence $\{\hat{u}^k\}$ is bounded, then all its cluster points lie in the optimal set U_*, and we have $f(\hat{u}^k) \to f_* > -\infty$, $\tau_k \to f_*$, $h(\hat{u}^k)_+ \to 0$, and $V_k \xrightarrow{K} 0$.*

(iv) *If $\{\hat{u}^k\}$ has a cluster point $\bar{u}$, then $\bar{u} \in U_*$, $h(\hat{u}^k)_+ \to 0$, and $\underline{\lim}_k \tau_k \geq \underline{\lim}_k f(\hat{u}^k) \geq f_* > -\infty$; moreover, if $K' \subset K$ is such that $\hat{u}^k \xrightarrow{K'} \bar{u}$, then $V_k \xrightarrow{K'} 0$.*

(v) *The sequence $\{\hat{u}^k\}$ has a cluster point if the set U_* is nonempty and bounded.*

(vi) *The sequence $\{\hat{u}^k\}$ is bounded if such is the set $U := \{u \in C : h(u) \leq 0\}$.*

(vii) *Suppose that $\bar{u} \in U_*$ and there exists an iteration index k' such that*

$$f(\bar{u}) \le \pi(\hat{u}^k; c_k + 1) \quad \text{for all } k \ge k', k \in K. \tag{3.13}$$

In particular, (3.13) *holds if $\hat{u}^{k'} \in U$ for some k', or $c_k \ge \bar{\mu} - 1$ for all $k \ge k', k \in K$. Further, suppose $\overline{\lim}_{k\in K} t_k < \infty$. Then the sequence $\{\hat{u}^k\}$ converges to a point in U_*.*

(viii) *Suppose that $\{\hat{u}^k\}$ is bounded, but we have only $\sum_{k\in K} t_k = \infty$ instead of $\inf_{k\in K} t_k \ge t_{\min}$. Then $\{\hat{u}^k\}$ has a cluster point in U_*. Moreover, assertion* (vii) *still holds.*

Proof. First, recalling the "exact" relations (2.32)–(2.33), note that $\epsilon_k \ge 0$ and

$$e_C(\cdot; \tau_k) \ge e_C(\hat{u}^k; \tau_k) + \langle p^k, \cdot - \hat{u}^k \rangle - \epsilon_k \quad \text{with} \quad e_C(\hat{u}^k; \tau_k) = h(\hat{u}^k)_+. \tag{3.14}$$

By Remark 2.7(vi), the descent test (2.30) ensures that $0 < h(\hat{u}^{k+1}) \le h(\hat{u}^k)$ for all k if $\bar{k} = \infty$, $f_* \le f(\hat{u}^{k+1}) \le f(\hat{u}^k)$, and $h(\hat{u}^k) \le 0$ for all $k \ge \bar{k}$ otherwise.

(i) Use $f_{\hat{u}}^{\infty} = \lim_k f(\hat{u}^k) = \lim_k \tau_k$ in Theorem 3.9 and the closedness of C, f, h.

(ii) Use the proof of Theorem 3.9 if $\bar{k} < \infty$ or Theorem 3.10(vi) otherwise.

(iii) First, suppose that $\bar{k} = \infty$; i.e., consider phase 1 with $h(\hat{u}^k) > 0$ for all k.

Let $\bar{u}$ be a cluster point of $\{\hat{u}^k\}$. Then $\bar{u} \in C$, since $\{\hat{u}^k\} \subset C$ and C is closed. Pick $K' \subset K$ such that $\hat{u}^k \xrightarrow{K'} \bar{u}$. Then $f(\hat{u}^k) \xrightarrow{K'} f(\bar{u})$, $h(\hat{u}^k) \xrightarrow{K'} h(\bar{u}) \ge 0$ (f, h are continuous on C). Since $\epsilon_k, |p^k| \xrightarrow{K} 0$ by (ii), Lemma 2.5(iv) yields $V_k \xrightarrow{K'} 0$. Let $\bar{\tau}$ be any cluster point of $\{\tau_k\}_{k\in K'}$. Pick $K'' \subset K'$ such that $\tau_k \xrightarrow{K''} \bar{\tau}$. We have $\bar{\tau} \ge f(\bar{u})$ ($\tau_k \ge f(\hat{u}^k)$) and $\bar{\tau} < \infty$; otherwise for large $k \in K''$, $\tau_k \ge f(\mathring{u}) - h(\mathring{u})$ would give $e(\mathring{u}; \tau_k) = h(\mathring{u}) < 0$ by (2.2) and (1.2), and by (3.14) with $\epsilon_k, |p^k| \xrightarrow{K} 0$,

$$0 > h(\mathring{u}) = e_C(\mathring{u}; \tau_k) \ge h(\hat{u}^k)_+ + \langle p^k, \mathring{u} - \hat{u}^k \rangle - \epsilon_k \xrightarrow{K''} h(\bar{u})_+ \ge 0,$$

a contradiction. Since e_C is continuous on $C \times \mathbb{R}$, letting $k \xrightarrow{K''} \infty$ in (3.14) gives $e_C(\cdot; \bar{\tau}) \ge e_C(\bar{u}; \bar{\tau})$, i.e., $0 \in \partial e_C(\bar{u}; \bar{\tau})$. Since $h(\bar{u}) \ge 0$ and $\bar{\tau} \ge f(\bar{u})$, $0 \in \partial e_C(\bar{u}; \bar{\tau})$ in (2.3) implies $\bar{\tau} = f(\bar{u})$ and $h(\bar{u}) = 0$ (otherwise for $h_C := h + i_C$, $0 \in \partial h_C(\bar{u})$ would give $\min_C h \ge 0$, contradicting (1.2)). Hence, $\bar{u} \in U_*$ by Lemma 2.2 (using $\bar{\tau} = \pi(\bar{u}; \bar{c})$ for any $\bar{c} \ge 0$) and $f(\bar{u}) = f_*$. Since $h(\bar{u}) = 0$ and $\{h(\hat{u}^k)\}$ is nonincreasing, we obtain that $h(\hat{u}^k) \to 0$.

By considering any convergent subsequences, we deduce that $V_k \xrightarrow{K} 0$ and that f_* is the unique cluster point of $\{\tau_k\}_{k\in K}$ and $\{f(\hat{u}^k)\}_{k\in K}$. Hence, $\lim_l \tau_{k(l)-1} = \lim_l f(\hat{u}^{k(l)-1}) = f_*$. Since $f(\hat{u}^{k(l)}) \le \tau_k \le \tau_{k(l+1)-1}$ for $k(l) \le k < k(l+1)$ by Steps 3, 4, and 7, we obtain $\lim_k f(\hat{u}^k) = \lim_k \tau_k = f_*$.

Finally, for the remaining case of $\bar{k} < \infty$, use the monotonicity of $\{\tau_k = f(\hat{u}^k)\}_{k\ge\bar{k}}$ and the relations $\bar{\tau} = f(\bar{u})$, $h(\bar{u}) \le 0$ in the second to last paragraph to get $0 \in \partial e_C(\bar{u}; \bar{\tau})$ and $\bar{u} \in U_*$ from Lemma 2.2; the rest follows as before.

(iv) Use the proof of (iii), getting $\underline{\lim}_k f(\hat{u}^k) \ge f_*$ from Lemma 3.3.

(v) If $\bar{k} < \infty$, the set $\{u \in C : f(u) \le f(\hat{u}^{\bar{k}}), h(u) \le 0\}$ is bounded (such is U_*) and contains $\{\hat{u}^k\}_{k\ge\bar{k}}$. Suppose that $\bar{k} = \infty$. By Theorem 3.10(vi), there is $K' \subset K$ such that $\overline{\lim}_{k\in K'} f(\hat{u}^k) \le f_*$. Hence, for infinitely many k, $\hat{u}^k$ lies in the set $\{u \in C : f(u) \le f_* + 1, h(u) \le h(u^1)_+\}$, which is bounded (such is U_*). Therefore, $\{\hat{u}^k\}$ has a cluster point.

(vi) The set $\{u \in C : h(u) \le h(u^1)_+\}$ is bounded (such is U) and contains $\{\hat{u}^k\}$.

(vii) If $\bar{k} < \infty$, then for $k \ge \bar{k}$, $\hat{u}^k \in U$ implies $f(\bar{u}) = f_* \le f(\hat{u}^k) = \pi(\hat{u}^k; c_k + 1)$; together with Lemma 2.3, this validates our claim below (3.13). Let $k \in K$, $k \ge k'$.

Since (3.13) implies $e_C(\bar{u};\tau_k) \le e_C(\hat{u}^k;\tau_k)$ by Lemma 2.3, (3.14) yields $\langle p^k, \bar{u}-\hat{u}^k\rangle \le \epsilon_k$. Then, using the facts that $\hat{u}^{k+1}-\hat{u}^k = -t_k p^k$ by (2.18) and $v_k = t_k|p^k|^2+\epsilon_k$ by (2.23), we get

$$\begin{aligned}|\hat{u}^{k+1}-\bar{u}|^2 &= |\hat{u}^k-\bar{u}|^2 + 2\langle \hat{u}^{k+1}-\hat{u}^k, \hat{u}^k-\bar{u}\rangle + |\hat{u}^{k+1}-\hat{u}^k|^2\\ &\le |\hat{u}^k-\bar{u}|^2 + 2t_k\epsilon_k + 2t_k^2|p^k|^2 = |\hat{u}^k-\bar{u}|^2 + 2t_k v_k.\end{aligned}$$

Therefore, since $\overline{\lim}_{k\in K} t_k < \infty$, $\sum_{k\in K} v_k < \infty$ by (ii), and $|\hat{u}^{k+1}-\bar{u}|^2 = |\hat{u}^k-\bar{u}|^2$ if $k \notin K$, we deduce from [Pol83, Lem. 2.2.2] that the sequence $\{|\hat{u}^k-\bar{u}|\}$ converges. Thus the sequence $\{\hat{u}^k\}$ is bounded, and using (iii) we may choose $\bar{u} \in U_*$ as a cluster point of $\{\hat{u}^k\}$, in which case the sequence $\{|\hat{u}^k-\bar{u}|\}$ must converge to zero, i.e., $\hat{u}^k \to \bar{u}$.

(viii) Argue as for (ii) to get $\sum_{k\in K} v_k < \infty$. Since $v_k = t_k|p^k|^2+\epsilon_k$ (cf. (2.23)) and $\epsilon_k \ge 0$, we have $\underline{\lim}_{k\in K}|p^k|^2 = 0$ (using $\sum_{k\in K} t_k = \infty$) and $\lim_{k\in K}\epsilon_k = 0$. Thus, there is $\bar{K} \subset K$ such that $\epsilon_k, |p^k| \xrightarrow{\bar{K}} 0$. Let $\bar{u}$ be a cluster point of $\{\hat{u}^k\}_{k\in\bar{K}}$. To see that $\bar{u} \in U_*$, replace K by $\bar{K}$ in the proof of (iii). Hence, this point $\bar{u}$ may be used in the final part of the proof of (vii). □

Remark 3.13. (i) The condition $\epsilon_{\max} = 0$ in Theorem 3.12 means that the linearizations are exact *and* Step 3 is inactive. If we drop this condition in Step 3, so that Step 3 ensures $v_k \ge \kappa_h h_{\hat{u}}^k$ when $h_{\hat{u}}^k > 0$ in the exact case as well, then for $\epsilon_{\max} = 0$, both Theorems 3.12 and 3.10 hold with $\epsilon_f = \epsilon_h = 0$ in the bounds of (3.1).

(ii) Condition (3.13) was used in [SaS05, Prop. 4.3(ii)] with $c_k \equiv 0$. Since in this case, $f_* = \inf_C \pi(\cdot, c_k+1)$ iff $\bar{\mu} \le 1$ (cf. section 2.1), we conclude that at phase 1 ($\bar{k} = \infty$) condition (3.13) with $c_k \equiv 0$ may be expected to hold only if $\bar{\mu} \le 1$. (Also see section 4.4.)

4. Modifications. In this section we consider several useful modifications.

4.1. Alternative descent tests. As in [Kiw06a, section 4.3], at Steps 4 and 5 we may replace the predicted decrease $v_k = t_k|p^k|^2+\epsilon_k$ (cf. (2.23)) by the smaller quantity $w_k := t_k|p^k|^2/2+\epsilon_k$. Then Lemma 2.5(ii) is replaced by the fact that

$$w_k \ge -\epsilon_k \quad\Longleftrightarrow\quad t_k|p^k|^2/4 \ge -\epsilon_k \quad\Longleftrightarrow\quad w_k \ge t_k|p^k|^2/4.$$

Hence, $w_k \ge -\epsilon_k$ at Step 5 implies $w_k \le v_k \le 3w_k$ and $v_k \ge -\epsilon_k$ for the bounds (2.25)–(2.26), whereas for Step 4, the bound (2.27) is replaced by the fact that

$$V_k < (4\epsilon_{\max}/t_k)^{1/2}(1+|\hat{u}^k|) \quad \text{if} \quad w_k < -\epsilon_k.$$

The preceding results extend easily (in the proof of Lemma 3.7, $e_{k+1}(u^{k+1}) > [h_{\hat{u}}^k]_+ - \kappa w_k$ implies $e_{k+1}(u^{k+1}) > [h_{\hat{u}}^k]_+ - \kappa v_k$, whereas in the proofs of Theorems 3.9 and 3.10(i), we have $\sum_{k\in K} v_k \le 3\sum_{k\in K} w_k < \infty$). We add that [SaS05, Alg. 3.1] uses w_k instead of v_k.

As in [Kiw85, p. 227], we may replace the descent test (2.30) by the two-part test

$$h_u^{k+1} \le h_{\hat{u}}^k - \kappa v_k \qquad \text{if } h_{\hat{u}}^k > 0, \tag{4.1a}$$

$$f_u^{k+1} \le f_{\hat{u}}^k - \kappa v_k \quad\text{and}\quad h_u^{k+1} \le 0 \quad \text{if } h_{\hat{u}}^k \le 0. \tag{4.1b}$$

Since (2.30) implies (4.1), the latter test may produce faster convergence. In particular, at phase 2 ($h_{\hat{u}}^k \le 0$) the additional requirement $h_u^{k+1} \le -\kappa v_k$ of (2.30) may

hinder the progress of $\{\hat{u}^k\}$ towards the boundary of the feasible set. The preceding convergence results are not affected (since if (4.1) fails at a null step, then so does (2.30), whereas the requirements of (4.1) suffice for descent steps).

In connection with (4.1b), we add that if $h_{\hat{u}}^1 \le 0$, i.e., the starting point is approximately feasible, then the objective linearizations need not be defined at infeasible points. Specifically, if $h_u^{k+1} > 0$ in (4.1b), then a null step must occur; so we may skip evaluating f_u^{k+1} and choose $J_f^{k+1} \supset \hat{J}_f^k$ at Step 6 (without requiring $J_f^{k+1} \ni k+1$). In the proof of Lemma 3.7, using $v_k = -\check{e}_k(u^{k+1})$ (cf. (2.10)) and replacing (3.10) by

$$e_{k+1}(\cdot) := \begin{cases} f_{k+1}(\cdot) - f_{\hat{u}}^k & \text{if } h_u^{k+1} \le 0, \\ h_{k+1}(\cdot) & \text{otherwise,} \end{cases} \tag{4.2}$$

we see that (4.1b) can be expressed as $e_{k+1}(u^{k+1}) \le -\kappa v_k$ or equivalently by (3.11); this suffices for the proof. Similarly, if $h_u^{k+1} \le 0$, then we may skip finding the subgradient g_h^{k+1} and choose $J_h^{k+1} \supset \hat{J}_h^k$ at Step 6 (omitting $\check{h}_k$ in (2.8) if $J_h^k = \emptyset$).

4.2. Linearization aggregation. To trade off storage and work per iteration for speed of convergence, one may replace selection with aggregation, so that only $\bar{m} \ge 4$ subgradients are stored. To this end, we note that the preceding results remain valid if, for each k, $\check{f}_{k+1}$ and $\check{h}_{k+1}$ are closed convex functions such that $0 \in \partial\phi_k(u^{k+1})$ implies (2.11)–(2.13) for k increased by 1, and

$$\max\{\bar{f}_k(u), f_{k+1}(u)\} \le \check{f}_{k+1}(u) \le f(u) \quad \text{for all } u \in C, \tag{4.3a}$$

$$\max\{\bar{h}_k(u), h_{k+1}(u)\} \le \check{h}_{k+1}(u) \le h(u) \quad \text{for all } u \in C. \tag{4.3b}$$

(This extends some ideas of [CoL93].) The max terms above are needed only after null steps in the proof of Lemma 3.7, $\bar{f}_k$ is not needed if $\nu_k = 0$, and $\bar{h}_k$ is not needed if $\nu_k = 1$. The aggregate linearizations may be treated like the oracle linearizations. Indeed, letting $f_{-j} := \bar{f}_j$, $h_{-j} := \bar{h}_j$ for $j = 1, \dots, k$, to ensure that $\bar{f}_k \le \check{f}_{k+1}$ and $\bar{h}_k \le \check{h}_{k+1}$, we may work with $J_f^{k+1}, J_h^{k+1} \subset \{-k, -k+1, \dots, k+1\}$ in (2.31), replacing the set $\hat{J}_f^k$ or $\hat{J}_h^k$ by $\{-k\}$ when $\hat{J}_f^k$ or $\hat{J}_h^k$ is "too large."

To illustrate, consider the following scheme with *minimal aggregation*. First, suppose $|J_f^k| + |J_h^k| = \bar{m}$. If $|\hat{J}_f^k| + |\hat{J}_h^k| \le \bar{m} - 2$, remove from J_f^k or J_h^k two indices in $J_f^k \setminus \hat{J}_f^k$ or $J_h^k \setminus \hat{J}_h^k$. If $|\hat{J}_f^k| + |\hat{J}_h^k| = \bar{m} - 1$, set $J_f^k := \hat{J}_f^k$, $J_h^k := \hat{J}_h^k$; if $|\hat{J}_h^k| \ge 2$, remove two indices from $\hat{J}_h^k$ and set $J_h^k := \hat{J}_h^k \cup \{-k\}$; otherwise, remove two indices from $\hat{J}_f^k$ and set $J_f^k := \hat{J}_f^k \cup \{-k\}$. If $|\hat{J}_f^k| + |\hat{J}_h^k| = \bar{m}$, remove four indices from $\hat{J}_f^k$ or $\hat{J}_h^k$, and set $J_f^k := \hat{J}_f^k \cup \{-k\}$, $J_h^k := \hat{J}_h^k \cup \{-k\}$. Next, suppose $|J_f^k| + |J_h^k| = \bar{m} - 1$. If $|\hat{J}_f^k| + |\hat{J}_h^k| = \bar{m} - 1$, proceed as in the second case above. If $|\hat{J}_f^k| + |\hat{J}_h^k| \le \bar{m} - 2$, remove from J_f^k or J_h^k one index in $J_f^k \setminus \hat{J}_f^k$ or $J_h^k \setminus \hat{J}_h^k$. At this stage, $|J_f^k| + |J_h^k| \le \bar{m} - 2$; so set $J_f^{k+1} := J_f^k \cup \{k+1\}$, $J_h^{k+1} := J_h^k \cup \{k+1\}$. This scheme employs aggregation only where needed; for $\bar{m} \ge m + 3$, it reduces to selection (cf. Remark 2.7(vii)).

In practice, without storing the points u^j for $j \ge 1$, we may use the representations

$$f_j(\cdot) = f_j(\hat{u}^k) + \langle \nabla f_j, \cdot - \hat{u}^k \rangle \quad \text{and} \quad h_j(\cdot) = h_j(\hat{u}^k) + \langle \nabla h_j, \cdot - \hat{u}^k \rangle,$$

since after a descent step, we can update the linearization values

$$f_j(\hat{u}^{k+1}) = f_j(\hat{u}^k) + \langle \nabla f_j, \hat{u}^{k+1} - \hat{u}^k \rangle \quad \text{for } j \in J_f^{k+1}, \tag{4.4a}$$

$$h_j(\hat{u}^{k+1}) = h_j(\hat{u}^k) + \langle \nabla h_j, \hat{u}^{k+1} - \hat{u}^k \rangle \quad \text{for } j \in J_h^{k+1}. \tag{4.4b}$$

Let us now consider *total aggregation*, in which only $\bar{m} \geq 2$ linearizations need be stored. Define e_1 by (3.10) with $k = 0$ and $\tau_0 := \tau_1$. Let $J_e^1 := \{1\}$. For $k \geq 1$, having linearizations $e_j(\cdot) \leq e(\cdot; \tau_k)$ for $j \in J_e^k$, replace $\check{e}_k$ in (2.8) by the "overall" model

$$\check{e}_k(\cdot) := \max_{j \in J_e^k} e_j(\cdot) \tag{4.5}$$

of $e(\cdot; \tau_k)$; thus we still have $\check{e}_k(\cdot) \leq e(\cdot; \tau_k)$ without maintaining separate models of f and h. Then the optimality condition $0 \in \partial \phi_k(u^{k+1})$ yields the existence of a subgradient $p_e^k \in \partial \check{e}_k(u^{k+1})$ such that p_e^k replaces $\nu_k p_f^k + (1 - \nu^k) p_h^k$ in (2.12) and (2.18). Consequently, using the aggregate linearization

$$\bar{e}_k(\cdot) := \check{e}_k(u^{k+1}) + \langle p_e^k, \cdot - u^{k+1} \rangle \leq \check{e}_k(\cdot) \leq e(\cdot; \tau_k) \tag{4.6}$$

and replacing the definition (2.17) of the linearization $\bar{e}_C^k$ and its expression (2.20) by

$$\bar{e}_C^k(\cdot) := \bar{e}_k(\cdot) + \bar{\imath}_C^k(\cdot) = \check{e}_k(u^{k+1}) + \langle p^k, \cdot - u^{k+1} \rangle \tag{4.7}$$

yields (2.21)–(2.22) and Lemma 2.5 as before. With e_{k+1} given by (3.10), for linearization selection we may use multipliers γ_j^k of the pieces e_j, $j \in J_e^k$, such that

$$(p_e^k, 1) = \sum_{j \in J_e^k} \gamma_j^k (\nabla e_j, 1),\ \gamma_j^k \geq 0,\ \gamma_j^k \big[\check{e}_k(u^{k+1}) - e_j(u^{k+1})\big] = 0,\ j \in J_e^k, \tag{4.8}$$

to choose the set $J_e^{k+1} \supset \hat{J}_e^k \cup \{k+1\}$ with $\hat{J}_e^k := \{j \in J_e^k : \gamma_j^k \neq 0\}$. For aggregation (cf. (4.3)), after a null step the next model $\check{e}_{k+1}$ should satisfy

$$\max\big\{\bar{e}_k(u), e_{k+1}(u)\big\} \leq \check{e}_{k+1}(u) \leq e(u; \tau_k) \quad \text{for all } u \in C, \tag{4.9}$$

and it suffices to choose $J_e^{k+1} \supset \{-k, k+1\}$ with $e_{-k} := \bar{e}_k$. Note that (4.6) and the minorization $e_{k+1}(\cdot) \leq e(\cdot; \tau_k)$ (cf. (3.10)) yield $\check{e}_{k+1}(\cdot) \leq e(\cdot; \tau_k)$. To ensure that $e(\cdot; \tau_k)$ is still minorized by each $e_j(\cdot) = e_j(\hat{u}^k) + \langle \nabla e_j, \cdot - \hat{u}^k \rangle$ after a descent step, since $e(\cdot; \tau_{k+1}) \geq e(\cdot; \tau_k) - (\tau_{k+1} - \tau_k)_+$ (cf. (2.2)), we may update

$$e_j(\hat{u}^{k+1}) := e_j(\hat{u}^k) + \langle \nabla e_j, \hat{u}^{k+1} - \hat{u}^k \rangle - (\tau_{k+1} - \tau_k)_+. \tag{4.10}$$

Similarly, when τ_k increases to τ_k', say, at Steps 3 or 4, the update $e_j(\hat{u}^k) := e_j(\hat{u}^k) - \tau_k' + \tau_k$ provides the minorization $e_j(\cdot) \leq e(\cdot; \tau_k')$.

Although total aggregation needs only $\bar{m} \geq 2$ linearizations, whereas separate aggregation described below (4.3) needs $\bar{m} \geq 4$, in practice this difference is immaterial, since larger values of $\bar{m}$ are required for faster convergence anyway. On the other hand, total aggregation has a serious drawback: its update (4.10), being based on a crude pessimistic estimate, tends to make the linearizations e_j lower than necessary when $\tau_{k+1} \neq \tau_k$. In contrast, separate aggregation is not sensitive to changes of τ_k.

Similar techniques can be applied to the *composite model*

$$\check{e}_k(\cdot) := \max \left\{ \max_{j \in J_f^k} f_j(\cdot) - \tau_k, \max_{j \in J_h^k} h_j(\cdot), \max_{j \in J_e^k} e_j(\cdot) \right\}. \tag{4.11}$$

For instance, (4.9) holds if $J_f^{k+1} \ni k+1$, $J_h^{k+1} \ni k+1$, $J_e^{k+1} \ni -k$, but many other choices are possible.

Remark 4.1. We add that [SaS05, Alg. 3.1] employs the model (4.11) with

$$J_f^k := \{\, j \in J^k : f_u^j - \tau_k \geq h_u^j \,\} \quad \text{and} \quad J_h^k := \{\, j \in J^k : f_u^j - \tau_k < h_u^j \,\} \tag{4.12}$$

for an additional "oracle" set $J^k \subset \{1, \ldots, k\}$; then J^k and J_e^k are reduced if necessary so that $2|J^k| + |J_e^k| \leq \bar{m} - 3$ for a given $\bar{m} \geq 3$, and $J^{k+1} := J^k \cup \{k+1\}$, $J_e^{k+1} := J_e^k \cup \{-k\}$. First, this scheme is quite unusual: although $|J^k|$ "original" linearizations of f and h are maintained ($2|J^k|$ in total), only half of them are selected via (4.12) for the model (4.11), thus reducing the QP size from $2|J^k| + |J_e^k|$ to $|J^k| + |J_e^k|$. (This selection is unnecessary in the sense that even for $J_f^k = J_h^k = J^k$, the model (4.11) still satisfies $\check{e}_k(\cdot) \leq e(\cdot, \tau_k)$.) Second, its storage requirement of $\bar{m} \geq 3$ places it between total aggregation and separate aggregation. Third, this scheme employs the update of (4.10) for $j \in J_e^k$.

4.3. Estimating Lagrange multipliers. Suppose that $f_* > -\infty$, so that the dual optimal set $M := \operatorname{Arg\,max}_{\mathbb{R}_+} q$ is nonempty (cf. section 2.1). For $\bar{\epsilon} \geq 0$, the set of *$\bar{\epsilon}$-optimal dual solutions* is defined by

$$M_{\bar{\epsilon}} := \{\, \mu \in \mathbb{R}_+ : q(\mu) \geq f_* - \bar{\epsilon} \,\}. \tag{4.13}$$

We now develop conditions under which the *Lagrange multiplier estimates*

$$\mu_k := (1 - \nu_k)/\nu_k \tag{4.14}$$

converge to the set $M_{\bar{\epsilon}}$ for a suitable $\bar{\epsilon} \geq 0$, where ν_k is the multiplier of (2.12)–(2.13).

Since $\nu_k \in [0, 1]$ by (2.13), (2.14)–(2.19) yield the sharper version of (2.22):

$$\nu_k \big[f(u) - \tau_k \big] + (1 - \nu_k) h(u) \geq \big[h_{\hat{u}}^k \big]_+ - V_k \big(1 + |u| \big) \quad \text{for all } u \in C. \tag{4.15}$$

If $\nu_k > 0$ (e.g., $V_k < -h(\mathring{u})/(1 + |\mathring{u}|)$), then (4.14) with $\mu_k \in \mathbb{R}_+$ and (4.15) give

$$f(u) + \mu_k h(u) \geq \tau_k - V_k \big(1 + |u| \big)/\nu_k \quad \text{for all } u \in C. \tag{4.16}$$

LEMMA 4.2. (i) *Suppose that $f_* > -\infty$. Let $K' \subset \mathbb{N}$ be such that $V_k \overset{K'}{\longrightarrow} 0$ and*

$$\varliminf_{k \in K'} \tau_k \geq f_* - \epsilon_f - \bar{\mu} \epsilon_h, \tag{4.17}$$

where $\bar{\mu} := \inf_{\mu \in M} \mu$ (cf. section 2.1). Then $\overline{\lim}_{k \in K'} \mu_k < \infty$ and $V_k/\nu_k \overset{K'}{\longrightarrow} 0$. Moreover, the sequence $\{\mu_k\}_{k \in K'}$ converges to the set $M_{\bar{\epsilon}}$ given by (4.13) for $\bar{\epsilon} := \epsilon_f + \bar{\mu} \epsilon_h$.

(ii) *If $f_* > -\infty$, then a set K' satisfying the requirements of* (i) *exists under the assumptions of Theorems* 3.8, 3.9, *or* 3.10 *or those of Theorem* 3.12 *if additionally either* $\inf\{k : h(\hat{u}^k) \leq 0\} < \infty$ *or* $|\hat{u}^k| \not\to \infty$ (*e.g., the optimal set U_* is nonempty and bounded*).

Proof. (i) By (4.17), $\tau_\infty := \varliminf_{k \in K'} \tau_k \geq f_* - \bar{\epsilon}$. If we had $\varliminf_{k \in K'} \nu_k = 0$, for $u = \mathring{u}$, (4.15) would yield in the limit $0 > h(\mathring{u}) \geq 0$, a contradiction. Hence, $\varliminf_{k \in K'} \nu_k > 0$, so that $V_k/\nu_k \overset{K'}{\longrightarrow} 0$ and $\overline{\lim}_{k \in K'} \mu_k < \infty$ by (4.14). Let μ_∞ be any cluster point of $\{\mu_k\}_{k \in K'}$; then $\mu_\infty \in \mathbb{R}_+$. Passing to the limit in (4.16) bounds the Lagrangian values as follows:

$$L(u; \mu_\infty) := f(u) + \mu_\infty h(u) \geq \tau_\infty \quad \text{for all } u \in C.$$

Hence, $q(\mu_\infty) \geq \tau_\infty \geq f_* - \bar{\epsilon}$ implies $\mu_\infty \in M_{\bar{\epsilon}}$ by (4.13). Since μ_∞ was an arbitrary cluster point of $\{\mu_k\}_{k \in K'} \subset \mathbb{R}_+ \cup \{\infty\}$ and $\overline{\lim}_{k \in K'} \mu_k < \infty$, the conclusion follows.

(ii) In Theorem 3.8, $\tau_k = f_{\hat{u}}^{\bar{k}}$ for all $k \geq \bar{k}$ (and we may take $K' = K$). In Theorem 3.9, $\tau_k \to f_{\hat{u}}^\infty \in [f_* - \epsilon_f - \bar{\mu} \epsilon_h, f_*]$ and $\varliminf_{k \in K} V_k = 0$. For the rest, see Theorems 3.10(ii, v) and 3.12(i, iv, v), noting that $|\hat{u}^k| \not\to \infty$ iff $\{\hat{u}^k\}$ has a cluster point. ∎

4.4. Updating the penalty coefficient in the exact case. We first show how to choose the penalty coefficient c_k by using the Lagrange multiplier estimate μ_k of (4.14) to ensure the "convergence" condition (3.13) of Theorem 3.12(vii).

LEMMA 4.3. *Under the assumptions of Theorem* 3.12*, suppose that* $|\hat{u}^k| \not\to \infty$. *Moreover, suppose that for all large* k*, after a descent step, Step* 7 *chooses* $c_{k+1} \geq \max\{\mu_k, c_k\}$ *if* $\mu_k < \infty$, $c_{k+1} \geq c_k$ *otherwise. Then there exists* k' *such that condition* (3.13) *holds for any* $\bar{u} \in U_*$.

Proof. By Theorem 3.12(iv), the assumptions of Lemma 4.2(i) hold for some $K' \subset K$, $\epsilon_f = \epsilon_h = \bar{\epsilon} = 0$; thus, $\{\mu_k\}_{k\in K'}$ converges to $M_0 = M$, and $\underline{\lim}_{k\in K'} \mu_k \geq \bar{\mu} := \inf_{\mu\in M} \mu$ implies $\mu_k \geq \bar{\mu} - 1$ for all large $k \in K'$. Hence, since $\{c_k\}$ is nondecreasing for large k, we have $c_k \geq \bar{\mu} - 1$ for all large k, and the conclusion follows from Theorem 3.12(vii). ∎

Remark 4.4. Variations on the strategy of Lemma 4.3 are possible. For instance, if $\{\hat{u}^k\}$ is bounded (e.g., U is bounded), Step 7 may choose $c_{k+1} \geq \mu_k$ after each descent step when $\mu_k < \infty$; this suffices for the proof of Lemma 4.3 with $K' = K$ by Theorem 3.12(iii).

We shall exploit the following basic property of the exact penalty function (2.1).

LEMMA 4.5. *If* $c \geq \bar{\mu}$, *then* $\pi(u; c) \geq f_* + (c - \bar{\mu})h(u)_+$ *for all* $u \in C$.

Proof. By (2.1), $\pi(u; c) = L(u; \bar{\mu}) + (c-\bar{\mu})h(u)_+ + \bar{\mu}[h(u)_+ - h(u)]$ for each $u \in C$, where $L(u; \bar{\mu}) \geq q(\bar{\mu}) = f_*$ (cf. section 2.1), $\bar{\mu} \geq 0$, and $h(u)_+ \geq h(u)$. ∎

For phase 1 in the exact case (when Step 3 is inactive), the main difficulty lies in ensuring $h(\hat{u}^k) \downarrow 0$. Complementing Theorem 3.12, we now show that it suffices if the penalty parameter c_k majorizes strictly the minimal Lagrange multiplier $\bar{\mu}$ asymptotically, and we give a specific update of c_k, based on a simple idea: increase the penalty coefficient if the constraint violation is large relative to the optimality measure (cf. [Kiw91]).

LEMMA 4.6. *Under the assumptions of Theorem* 3.12*, suppose that* $h(\hat{u}^k) > 0$ *for all* k. *Then we have the following statements*:

(i) *There is* $K' \subset K$ *such that* $V_k \xrightarrow{K'} 0$ *and* $\overline{\lim}_{k\in K'} f(\hat{u}^k) \leq \overline{\lim}_{k\in K'} \tau_k \leq f_*$.

(ii) *If* $c_\infty := \underline{\lim}_k c_k > \bar{\mu}$, *then* $h(\hat{u}^k) \downarrow 0$.

(iii) *Suppose that for all large* k*, after a descent step, Step* 7 *chooses* $c_{k+1} \geq 2c_k$ *if* $h(\hat{u}^{k+1}) > V_k$, $c_{k+1} \geq c_k$ *otherwise,* $c_{k+1} > 0$ *when* $h(\hat{u}^{k+1}) > 0$. *If* $f_* > -\infty$, *then* $h(\hat{u}^k) \downarrow 0$.

(iv) *If* $h(\hat{u}^k) \downarrow 0$, *then* $\underline{\lim}_k \tau_k \geq \underline{\lim}_k f(\hat{u}^k) \geq f_*$, *and* $f(\hat{u}^k) \xrightarrow{K'} f_*$ *in* (i) *above.*

Proof. (i) This follows from Theorem 3.10(vi).

(ii) By (i) and Lemma 4.5, $f_* \geq \underline{\lim}_k \tau_k \geq f_* + (c_\infty - \bar{\mu}) \underline{\lim}_k h(\hat{u}^k)_+$ with $c_\infty > \bar{\mu}$ yields $\underline{\lim}_k h(\hat{u}^k)_+ = 0$. Hence, $h(\hat{u}^k) \downarrow 0$, using $0 < h(\hat{u}^{k+1}) \leq h(\hat{u}^k)$ by (2.34a).

(iii) If $c_\infty := \lim_k c_k < \infty$, then $h(\hat{u}^{k+1}) \leq V_k$ for all large $k \in K$; so by (i), $V_k \xrightarrow{K'} 0$ yields $h(\hat{u}^k) \downarrow 0$. Otherwise, $c_\infty = \infty > \bar{\mu}$ (from $f_* > -\infty$), and (ii) applies.

(iv) Invoke Lemma 3.3 with $\epsilon_f = \epsilon_h = 0$, and use the fact that $\tau_k \geq f(\hat{u}^k)$. ∎

5. Column generation for LP problems. In this section we consider the following primal-dual pair of LP problems:

$$\min\ c\lambda \quad \text{s.t.} \quad A\lambda \geq b,\ \lambda \geq 0, \tag{5.1}$$

$$\max\ ub \quad \text{s.t.} \quad uA \leq c,\ u \geq 0, \tag{5.2}$$

where $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$. We assume that $c > 0$. Let A_i denote column i of A for $i \in I := \{1\colon n\}$. When the number of columns is huge, problems (5.1)–(5.2)

may be solved by column generation, provided that for each $u \geq 0$, one can solve the *column generation subproblem* of finding $i_u \in \operatorname{Arg\,max}_{i \in I}(uA_i - c_i)$. We show that this subproblem may be solved inexactly when our method is applied to the dual problem (5.2) formulated as (1.1) and that approximate solutions to (5.1) can be recovered at no extra cost.

To ease subsequent notation, let us rewrite the LP problems (5.1)–(5.2) as follows:

$$\max\ \psi_0(\lambda) := -c\lambda \quad \text{s.t.} \quad \psi(\lambda) := A\lambda - b \geq 0,\ \lambda \in \mathbb{R}^n_+, \tag{5.3}$$

$$\min\ f(u) := -ub \quad \text{s.t.} \quad uA \leq c,\ u \in \mathbb{R}^m_+. \tag{5.4}$$

We regard the dual problem (5.4) as (1.1) with $C := \mathbb{R}^m_+$ and the constraint function

$$h(\cdot) := \max_{i \in I}\bigl(\langle A_i, \cdot\rangle - c_i\bigr). \tag{5.5}$$

Since $c > 0$, $\mathring{u} := 0$ may serve as the Slater point. For our method applied to (1.1), we assume that f is evaluated exactly (i.e., $\epsilon_f = 0$ and $f_k = f$), whereas the approximate linearization condition (2.4b) boils down to finding an index $i_k \in I$ such that

$$h_k(\cdot) = \langle A_{i_k}, \cdot\rangle - c_{i_k} \quad \text{with} \quad h_k(u^k) \geq h(u^k) - \epsilon_h. \tag{5.6}$$

By duality, f_* is the common optimal value of (5.3) and (5.4). In view of Lemma 4.2, we assume that $f_* > -\infty$ and let $K' \subset \mathbb{N}$ be the set such that $V_k \xrightarrow{K'} 0$ and (4.17) holds; then $\nu_k > 0$ and $\mu_k := (1 - \nu_k)/\nu_k < \infty$ for large $k \in K'$. We shall show that the corresponding subsequence of the multipliers $\{\mu_k \beta^k_j\}_{j \in J^k_h}$ of (2.28b) solves the primal problem (5.3) approximately; thus, below we consider only $k \in K'$ such that $\nu_k > 0$.

The multipliers $\{\mu_k \beta^k_j\}_{j \in J^k_h}$ define an *approximate primal solution* $\hat{\lambda}^k \in \mathbb{R}^n_+$ via

$$\hat{\lambda}^k_i := \mu_k \sum_{j \in J^k_h : i_j = i} \beta^k_j \quad \text{for each } i \in I.$$

Let $\underline{1} := (1, \dots, 1) \in \mathbb{R}^n$. In this notation, using the form (5.6) of the linearizations h_j in (2.28b) and the fact that $\mu_k \check{h}_k(u^{k+1}) = \mu_k \check{e}_k(u^{k+1})$ (cf. (2.13)) yields the relations

$$\mu_k p^k_h = A\hat{\lambda}^k, \quad \mu_k = \underline{1}\hat{\lambda}^k, \quad \hat{\lambda}^k \geq 0, \quad (u^{k+1}A - c)\hat{\lambda}^k = \mu_k \check{e}_k(u^{k+1}). \tag{5.7}$$

We first derive useful expressions for the primal function values $\psi_0(\hat{\lambda}^k)$ and $\psi(\hat{\lambda}^k)$.

LEMMA 5.1. $\psi_0(\hat{\lambda}^k) = \tau_k + ([h^k_{\hat{u}}]_+ - \epsilon_k - \langle p^k, \hat{u}^k\rangle)/\nu_k$, $\psi(\hat{\lambda}^k) = (p^k - p^k_C)/\nu_k \geq p^k/\nu_k$.

Proof. Since $p^k_f = \nabla f = -b$ (cf. (2.11), (5.4)), $\mu_k p^k_h = A\hat{\lambda}^k$ by (5.7), and $\nu_k \mu_k = 1 - \nu_k$ by (4.14), the definitions of $\psi(\lambda)$ in (5.3) and of p^k in (2.18) give

$$\nu_k \psi(\hat{\lambda}^k) = \nu_k\bigl(A\hat{\lambda}^k - b\bigr) = \nu_k p^k_f + (1 - \nu_k)p^k_h = p^k - p^k_C,$$

where $p^k_C \in \partial i_{\mathbb{R}^m_+}(u^{k+1})$ implies $p^k_C \leq 0$ and $\langle p^k_C, u^{k+1}\rangle = 0$. Next, by (5.7) and (2.18),

$$\begin{aligned} \nu_k c\hat{\lambda}^k + (1 - \nu_k)\check{e}_k(u^{k+1}) &= \langle \nu_k \mu_k p^k_h, u^{k+1}\rangle \\ &= \langle (1 - \nu_k)p^k_h + p^k_C, u^{k+1}\rangle = \langle p^k - \nu_k p^k_f, u^{k+1}\rangle, \end{aligned}$$

where $\nu_k\langle p_f^k, u^{k+1}\rangle = \nu_k \check{f}_k(u^{k+1}) = \nu_k \check{e}_k(u^{k+1}) + \nu_k\tau_k$ by (2.13). Hence,

$$-\nu_k c\hat{\lambda}^k - \nu_k\tau_k = \check{e}_k(u^{k+1}) - \langle p^k, u^{k+1}\rangle = \bar{e}_C^k(0) = \big[h_{\hat{u}}^k\big]_+ - \langle p^k, \hat{u}^k\rangle - \epsilon_k,$$

where we have used (2.20)–(2.21). Dividing by ν_k gives the required expression of $\psi_0(\hat{\lambda}^k) := -c\hat{\lambda}^k$; for $\psi(\hat{\lambda}^k)$, see the first displayed equality above. ∎

In terms of the optimality measure V_k of (2.19), the bounds of Lemma 5.1 imply

$$\hat{\lambda}^k \geq 0 \quad \text{with} \quad \psi_0(\hat{\lambda}^k) \geq \tau_k - V_k/\nu_k, \quad \psi_i(\hat{\lambda}^k) \geq -V_k/\nu_k, \quad i = 1\colon m. \tag{5.8}$$

We now show that $\{\hat{\lambda}^k\}_{k\in K'}$ converges to the set of $\bar{\epsilon}$-*optimal primal solutions*

$$\Lambda_{\bar{\epsilon}} := \big\{\, \lambda \in \mathbb{R}^n_+ : \psi_0(\lambda) \geq f_* - \bar{\epsilon}, \psi(\lambda) \geq 0 \,\big\}, \tag{5.9}$$

where $\bar{\epsilon} := \bar{\mu}\epsilon_h$, with $\bar{\mu}$ being the minimal Lagrange multiplier of (1.1); in our context, we may as well take (a possibly larger) $\bar{\mu} := \underline{1}\bar{\lambda}$ for any primal solution $\bar{\lambda}$ of (5.3).

THEOREM 5.2. *Suppose that $f_* > -\infty$. Let $K' \subset \mathbb{N}$ be such that $V_k \xrightarrow{K'} 0$ and* (4.17) *holds* (*see Lemma* 4.2(ii) *for sufficient conditions*). *Then we have the following statements*:

(i) *The sequence $\{\hat{\lambda}^k\}_{k\in K'}$ is bounded and all its cluster points lie in $\mathbb{R}^n_+$.*

(ii) *Let $\hat{\lambda}^\infty$ be a cluster point of $\{\hat{\lambda}^k\}_{k\in K'}$. Then $\hat{\lambda}^\infty \in \Lambda_{\bar{\epsilon}}$.*

(iii) $d_{\Lambda_{\bar{\epsilon}}}(\hat{\lambda}^k) := \inf_{\lambda\in\Lambda_{\bar{\epsilon}}} |\hat{\lambda}^k - \lambda| \xrightarrow{K'} 0$.

Proof. By Lemma 4.2, $\overline{\lim}_{k\in K'} \mu_k < \infty$ and $V_k/\nu_k \xrightarrow{K'} 0$. Since $\underline{\lim}_{k\in K'} \tau_k \geq f_* - \bar{\epsilon}$ by (4.17), (5.8) yields $\underline{\lim}_{k\in K'} \psi_0(\hat{\lambda}^k) \geq f_* - \bar{\epsilon}$ and $\underline{\lim}_{k\in K'} \min_{i=1}^n \psi_i(\hat{\lambda}^k) \geq 0$.

(i) This follows from $\overline{\lim}_{k\in K'} \underline{1}\hat{\lambda}^k = \overline{\lim}_{k\in K'} \mu_k < \infty$ (cf. (5.7)) and $\hat{\lambda}^k \geq 0$.

(ii) We have $\hat{\lambda}^\infty \geq 0$, $\psi_0(\hat{\lambda}^\infty) \geq f_* - \bar{\epsilon}$, and $\psi(\hat{\lambda}^\infty) \geq 0$ by continuity of ψ_0 and ψ.

(iii) Use (i), (ii), and the continuity of the distance function $d_{\Lambda_{\bar{\epsilon}}}$. ∎

Remark 5.3. (i) By Remark 3.11(ii), we may use $\bar{\epsilon} := \bar{\mu}\epsilon_h^\infty$ for Theorem 5.2.

(ii) By Lemma 3.1(iii) and the proof of Theorem 5.2, if an infinite loop between Steps 1 and 4 occurs, then $V_k \to 0$ yields $d_{\Lambda_{\bar{\epsilon}}}(\hat{\lambda}^k) \to 0$. Similarly, if Step 2 terminates with $V_k = 0$, then $\hat{\lambda}^k \in \Lambda_{\bar{\epsilon}}$. In both cases, we may take $\bar{\epsilon} := \bar{\mu}\epsilon_h^{k(l)}$ by Remark 3.11(ii).

(iii) Given two tolerances $\epsilon_{\rm F}, \epsilon_{\rm tol} > 0$, the method may stop if $h_{\hat{u}}^k \leq \epsilon_{\rm F}$,

$$\psi_0(\hat{\lambda}^k) \geq f(\hat{u}^k) - \epsilon_{\rm tol} \quad \text{and} \quad \psi_i(\hat{\lambda}^k) \geq -\epsilon_{\rm tol}, \quad i = 1\colon m.$$

Then $\psi_0(\hat{\lambda}^k) \geq f_* - \bar{\mu}(\epsilon_h + \epsilon_{\rm F}) - \epsilon_{\rm tol}$ from $f(\hat{u}^k) \geq f_* - \bar{\mu}(\epsilon_h + \epsilon_{\rm F})$; so $\hat{\lambda}^k$ is an approximate solution of (5.3). This stopping criterion will be met when $V_k/\nu_k \leq \epsilon_{\rm tol}$.

We add that our numerical experiments (to be reported elsewhere) on the test problems of [Kiw05, KiL07, SaS05] indicate that our method is quite sensitive to constraint scaling; yet, with proper scaling, it can perform quite well.

Acknowledgments. I would like to thank the Associate Editor, the two anonymous referees, and Claude Lemaréchal for helpful comments.

REFERENCES

[Ber99] D. P. BERTSEKAS, *Nonlinear Programming*, 2nd ed., Athena Scientific, Belmont, MA, 1999.

[BLM+07] O. BRIANT, C. LEMARÉCHAL, PH. MEURDESOIF, S. MICHEL, N. PERROT, AND F. VANDERBECK, *Comparison of bundle and classical column generation*, Math. Program., to appear.

[BTN05] A. Ben-Tal and A. Nemirovski, *Non-euclidean restricted memory level method for large-scale convex optimization*, Math. Program., 102 (2005), pp. 407–456.

[CoL93] R. Correa and C. Lemaréchal, *Convergence of some algorithms for convex minimization*, Math. Program., 62 (1993), pp. 261–275.

[Fáb00] C. I. Fábián, *Bundle-type methods for inexact data*, CEJOR Cent. Eur. J. Oper. Res., 8 (2000), pp. 35–55.

[FlL99] R. Fletcher and S. Leyffer, *A Bundle Filter Method for Nonsmooth Nonlinear Optimization*, Numerical Analysis report NA/195, Department of Math., University of Dundee, Dundee, Scotland, 1999.

[HUL93] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms*, Springer-Verlag, Berlin, 1993.

[KRSS07] E. Karas, A. Ribeiro, C. Sagastizábal, and M. Solodov, *A bundle–filter method for nonsmooth convex constrained optimization*, Math. Program., to appear.

[Kiw85] K. C. Kiwiel, *Methods of Descent for Nondifferentiable Optimization*, Lecture Notes in Math. 1133, Springer-Verlag, Berlin, 1985.

[Kiw87] K. C. Kiwiel, *A constraint linearization method for nondifferentiable convex minimization*, Numer. Math., 51 (1987), pp. 395–414.

[Kiw90] K. C. Kiwiel, *Proximity control in bundle methods for convex nondifferentiable minimization*, Math. Program., 46 (1990), pp. 105–122.

[Kiw91] K. C. Kiwiel, *Exact penalty functions in proximal bundle methods for constrained convex nondifferentiable minimization*, Math. Program., 52 (1991), pp. 285–302.

[Kiw94] K. C. Kiwiel, *A Cholesky dual method for proximal piecewise linear programming*, Numer. Math., 68 (1994), pp. 325–340.

[Kiw95] K. C. Kiwiel, *Proximal level bundle methods for convex nondifferentiable optimization, saddle-point problems and variational inequalities*, Math. Program., 69 (1995), pp. 89–109.

[Kiw05] K. C. Kiwiel, *An Inexact Bundle Approach to Cutting-Stock Problems*, Tech. report, Systems Research Institute, Warsaw, Poland, 2005, revised 2006.

[Kiw06a] K. C. Kiwiel, *A proximal bundle method with approximate subgradient linearizations*, SIAM J. Optim., 16 (2006), pp. 1007–1023.

[Kiw06b] K. C. Kiwiel, *A proximal-projection bundle method for Lagrangian relaxation, including semidefinite programming*, SIAM J. Optim., 17 (2006), pp. 1015–1034.

[Kiw07a] K. C. Kiwiel, *An Inexact Filter Bundle Method for Nonsmooth Convex Optimization*, Tech. report, Systems Research Institute, Warsaw, Poland, 2007, in preparation.

[Kiw07b] K. C. Kiwiel, *A Penalty Bundle Method with Inexact Subgradient Linearizations*, Tech. report, Systems Research Institute, Warsaw, Poland, 2007, in preparation.

[KiL07] K. C. Kiwiel and C. Lemaréchal, *An inexact bundle variant suited to column generation*, Math. Program., to appear.

[LNN95] C. Lemaréchal, A. S. Nemirovskii, and Yu. E. Nesterov, *New variants of bundle methods*, Math. Program., 69 (1995), pp. 111–147.

[LeS97] C. Lemaréchal and C. Sagastizábal, *Variable metric bundle methods: From conceptual to implementable forms*, Math. Program., 76 (1997), pp. 393–410.

[LüD05] M. E. Lübbecke and J. Desrosiers, *Selected topics in column generation*, Oper. Res., 53 (2005), pp. 1007–1023.

[Pol83] B. T. Polyak, *Introduction to Optimization*, Nauka, Moscow, 1983 (in Russian); Optimization Software Inc., New York, 1987 (in English).

[Roc70] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[SaS05] C. Sagastizábal and M. Solodov, *An infeasible bundle method for nonsmooth convex constrained optimization without a penalty function or a filter*, SIAM J. Optim., 16 (2005), pp. 146–169.

[Sav97] M. W. P. Savelsbergh, *A branch-and-price algorithm for the generalized assignment problem*, Oper. Res., 45 (1997), pp. 831–841.

SIAM J. OPTIM.
Vol. 18, No. 4, pp. 1490–1500

IMPROVED APPROXIMATION ALGORITHMS FOR WEIGHTED HYPERGRAPH EMBEDDING IN A CYCLE*

HANN-JANG HO† AND SINGLING LEE‡

Abstract. We consider the problem of embedding weighted hyperedges of a hypergraph as paths in a cycle on the same number of vertices, such that the maximum congestion of any physical link of the cycle is minimized. The problem, called weighted hypergraph embedding in a cycle (WHEC), is known to be NP-complete even when each hyperedge is unweighted or each weighted hyperedge contains exactly two vertices. In this paper, we propose an improved rounding algorithm for the WHEC problem to provide a solution with an approximation bound of $1.5(opt + w_{max})$, where opt represents the optimal value of the problem and w_{max} denotes the largest weight of hyperedges. For any fixed $\varepsilon > 0$, we also present a polynomial time algorithm to provide an embedding whose congestion is at most $(1.5 + \varepsilon)$ times the optimum. This improves previous results for the general WHEC problem.

Key words. weighted hypergraph embedding in a cycle, linear programming, approximation algorithm, hypergraph, NP-complete

AMS subject classifications. 05C65, 68W25, 90C05, 90C27

DOI. 10.1137/050631951

1. Introduction.

The problem of weighted hypergraph embedding in a cycle (WHEC) is to embed m weighted hyperedges of a hypergraph as paths in an n-vertex cycle, such that the maximum congestion of any physical link in the cycle is minimized. Note that each vertex of the hypergraph actually is the same as the vertex in the cycle; i.e., the problem of mapping vertices between hypergraph and cycle is not considered here. The WHEC problem is NP-complete even when each hyperedge is unweighted or each weighted hyperedge contains exactly two vertices [1].

In the first special case when each hyperedge is unweighted, the unweighted version of the WHEC problem has many applications in electronic design automation [2, 3]. Several studies related to the unweighted WHEC problem have been performed. Frank et al. [4] and Gonzalez and Lee [2] proposed linear time algorithms to solve the problem when all hyperedges contain exactly only two vertices. Ganley and Cohoon [5] showed that the unweighted WHEC problem is NP-complete in general but solvable in polynomial time if the maximum congestion is bounded by a constant. Gonzalez [6] proposed two improved approximation algorithms that both generate solutions with maximum congestion at most two times the optimum. Carpenter et al. [7] provided a linear time approximation algorithm which routes the hyperedges in the clockwise direction starting from the lowest numbered vertex to the highest numbered vertex. This algorithm is also guaranteed to find a solution whose value is at most twice the optimal value. Recently, Gu and Wang presented an algorithm to solve the unweighted WHEC problem with the performance ratio 1.8 by a

*Received by the editors May 20, 2005; accepted for publication (in revised form) July 16, 2007; published electronically January 16, 2008. This work was supported in part by the Taiwan NSC under grants 94-2213-E-274-004 and 93-2213-E-194-038.

http://www.siam.org/journals/siopt/18-4/63195.html

†Corresponding author. Department of Computer Science and Information Engineering, Wufeng Institute of Technology, 117, Sec. 2, Jianguo Rd., Minsyong, Chiayi 621, Taiwan, R.O.C. (hjho@cs.ccu.edu.tw).

‡Department of Computer Science and Information Engineering, National Chung-Cheng University, 168 University Rd., Minsyong, Chiayi 621, Taiwan, R.O.C. (singling@ccu.edu.tw).

reembedding technique [8]. Lee and Ho presented an LP-based algorithm to solve the problem with the performance ratio 1.5 [9]. Deng and Li proposed a polynomial time approximation scheme (PTAS) for the unweighted WHEC problem by a randomized rounding approach [10].

In the second special case when each weighted hyperedge contains exactly two vertices, the graph version of the WHEC problem is equivalent to the ring loading problem without demand splitting. When each demand must be entirely routed in either of the two directions, the ring loading problem is NP-complete [11]. Cosares and Saniee presented a polynomial time algorithm that approximates the optimal solution value to within a multiplicative factor of two [11]. Schrijver, Seymour, and Winkler developed an efficient algorithm to generate a solution that exceeds the optimum by at most an additive term of 1.5 times the maximum weight [12]. Khanna proposed a PTAS that computes a solution with at most $(1+\varepsilon)$ times the optimum for any fixed $\varepsilon > 0$ [13].

The WHEC problem has many applications, including minimizing communication congestions in computer networks and parallel computations. These applications focus on minimizing the maximum congestions of multicast transmissions over any physical communication link. For example, in computer networks, the hypergraph embedding problem is equivalent to a multicast congestion problem [16, 17, 18] when each hyperedge is formed as a multicast tree (i.e., Steiner tree). Several studies related to the multicast congestion problem on more general graphs have been performed. Vempala and Vöcking [16] presented a randomized rounding algorithm for approximating multicast congestion to within $O(\log n)$ times the optimum. Carr and Vempala [17] proposed a randomized asymptotic algorithm for the multicast congestion problem to find a solution of congestion $O(opt + \log n)$, where *opt* is the optimal value of the maximum congestion. Jansen and Zhang [18] presented an improved approximation algorithm to overcome the difficulties of an exponential number of variables.

We are concerned with the general WHEC problem. In [1], an LP-based rounding algorithm and a linear time algorithm had been proposed to find a solution with maximum congestion at most two times the optimum. In this paper, we propose an improved (2/3)-rounding algorithm to provide a solution with an approximation bound of $1.5(opt + w_{max})$, where *opt* represents the optimal value of the problem and w_{max} denotes the largest weight of hyperedges. For any fixed $\varepsilon > 0$, we also present a PTAS to provide an embedding whose congestion is at most $(1.5+\varepsilon)$ times the optimum. This improves the previous approximation results [1] for the general WHEC problem.

2. Notation and problem definition. The WHEC problem is to embed the weighted hyperedges of a hypergraph as the paths in a cycle, such that the maximum congestion of any physical link in the cycle is minimized. We denote the cycle as $C = (V, E_c)$, where $V = \{1, 2, \ldots, n\}$ is a set of vertices and $E_c = \{(i, i+1) | 1 \leq i \leq n-1\} \bigcup \{(n, 1)\}$ is a set of physical links. The vertices of the cycle are labelled clockwise by 1 through n, and each edge in the cycle is referred to as an undirected link. A hypergraph $H = (V, E_h)$ with m weighted hyperedges is defined over the same set of vertices V, where $E_h = \{h_1, h_2, \ldots, h_m\}$ is a set of hyperedges. The hyperedge h_i consists of $|h_i|$ vertices with a nonnegative weight w_i for interconnecting these vertices. In particular, these interconnected vertices of hyperedge h_i are represented as an ordered sequence $(v_1^i, v_2^i, \ldots, v_{|h_i|}^i)$, i.e., $v_1^i \leq v_2^i \leq \cdots \leq v_{|h_i|}^i$. For example, a hypergraph $H = \{h_1, h_2, h_3\}$ is embedded in an 8-vertex cycle as shown in Figure 1.

We denote the jth adjacent path of hyperedge h_i as $p(i, j)$, where $p(i, j)$ is a

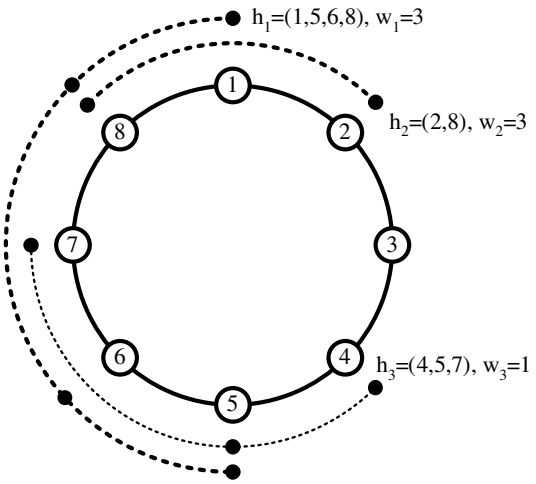


FIG. 1. *A weighted hypergraph* $H = \{h_1, h_2, h_3\}$ *is embedded in an* 8*-vertex cycle. The link* $(8, 1)$ *has the maximum congestion of* 6 *that is an optimal solution.*

clockwise connecting path between v_j^i and v_{j+1}^i in the cycle C, and the last adjacent path $p(i, |h_i|)$ of the hyperedge h_i is connected between $v_{|h_i|}^i$ and v_1^i. Obviously, vertices $v_1^i, v_2^i, \ldots, v_{|h_i|}^i$ in the hyperedge h_i must be connected by at least $|h_i| - 1$ adjacent paths to ensure the connectivity. We define an assignment of adjacent paths to embed the hypergraph H in the cycle C as a set of binary variables $Y = \{y_{p(i,j)} | \forall i \in \{1, 2, \ldots, m\}, j \in \{1, 2, \ldots, |h_i|\}\}$, where $y_{p(i,j)} = 1$ if an adjacent path $p(i, j)$ is embedded in the cycle C, and $y_{p(i,j)} = 0$ otherwise. Therefore, a feasible assignment of adjacent paths can be expressed as $\sum_{1 \le j \le |h_i|} y_{p(i,j)} \ge |h_i| - 1 \; \forall h_i \in E_h$. Moreover, we denote $P(e)$ as a set of adjacent paths that pass through the link $e \in E_c$. Therefore, the WHEC problem can be formally defined as follows.

DEFINITION 2.1. *The problem of minimizing the maximum congestion for WHEC is as follows*: *Given a cycle* $C = (V, E_c)$ *and a weighted hypergraph* $H = (V, E_h)$, *find a feasible assignment* Y *of adjacent paths to embed the hypergraph* H *in the cycle* C *such that the maximum congestion* φ *over any link in* E_c *is minimized. Note that each hyperedge* h_i *is associated with a nonnegative weight* w_i *and the maximum link congestion of an assignment* Y *can be expressed as* $\varphi = \max_{e \in E_c} \{\sum_{p(i,j) \in P(e)} y_{p(i,j)} w_i\}$, *where* $y_{p(i,j)} = 1$ *if the* $p(i, j)$ *is embedded in the cycle* C, *and* $y_{p(i,j)} = 0$ *otherwise.*

3. Mixed integer linear programming. The WHEC problem can be modelled as a mixed integer linear programming (MILP) formulation. We define a set of binary variables $Y = \{y_{p(i,j)} | \forall i \in \{1, 2, \ldots, m\}, j \in \{1, 2, \ldots, |h_i|\}\}$ to represent the assignment of adjacent paths to embed the hypergraph H in the cycle C, where $y_{p(i,j)} = 1$ if an adjacent path $p(i, j)$ is embedded in the cycle C, and $y_{p(i,j)} = 0$ otherwise. Each hyperedge h_i is associated with a nonnegative weight w_i. Our objective is to minimize the maximum congestion φ over all links in E_c. We formulate an MILP model to solve the problem as follows:

$\Phi(Y)$: Minimize φ

Subject to

$$\sum_{1\le j\le |h_i|} y_{p(i,j)} \ge |h_i| - 1 \quad \forall h_i \in E_h, \quad i \in \{1,2,\ldots,m\} \quad \text{(connectivity constraints)},$$

$$\sum_{p(i,j)\in P(e)} y_{p(i,j)} w_i \le \varphi \quad \forall e \in E_c \quad \text{(capacity constraints)},$$

$$y_{p(i,j)} \in \{0,1\} \quad \forall i \in \{1,2,\ldots,m\}, \quad j \in \{1,2,\ldots,|h_i|\} \quad \text{(binary variables)}.$$

There are two classes of constraints in this problem. First, the connectivity constraints mean that each hyperedge $h_i \in E_h$ must be connected by at least $|h_i| - 1$ adjacent paths. Second, the capacity constraints ensure that each physical link $e \in E_c$ can only be embedded with hyperedges at most the maximum congestion φ. Finally, the objective function $\Phi(Y)$ of the MILP is to minimize the maximum congestion φ. Note that φ is a continuous variable since the binary variable $y_{p(i,j)}$ is multiplied by a weight w_i which is a nonnegative real constant.

Suppose that we relax the integer constraints of Y and require only that $0 \le y_{p(i,j)} \le 1$ $\forall i,j$; this linear program is called the LP relaxation of the MILP formulation, and the LP solution immediately provides a lower bound on the minimum congestion. On the other hand, finding a tight upper bound for the MILP formulation presents a difficult challenge. However, the LP-rounding technique can commonly be applied to find an upper bound and generate an approximate solution. Let $\varphi^* = \Phi(Y^*)$ and $\varphi^L = \Phi(Y^L)$ be the optimal solutions of MILP formulation and LP relaxation, respectively. If the solution for the LP-rounding is φ^R, then we have $\varphi^L \le \varphi^* \le \varphi^R$.

The next lemma states that the value of the kth smallest variable in $\{y^L_{p(i,j)} \mid 1 \le j \le |h_i|\}$ is at least $(k-1)/k$ for every hyperedge $h_i \in E_h$. This implies that the value of the second smallest variable is at least $1/2$. Therefore, a $(1/2)$-rounding approach can be applied to generate an approximate solution $\varphi^R = \Phi(Y^R)$, where we assign $y^R_{p(i,j)} = 1$ if $y^L_{p(i,j)} \ge 1/2$, and $y^R_{p(i,j)} = 0$ otherwise. The $(1/2)$-rounding approach always produce a 2-approximated solution for the WHEC problem [1].

LEMMA 3.1. *For every hyperedge $h_i \in E_h$, $|h_i| \ge k$, $i \in \{1,2,\ldots,m\}$, the value of the kth smallest variable in $\{y^L_{p(i,j)} \mid 1 \le j \le |h_i|\}$ is at least $(k-1)/k$.*

Proof. From connectivity constraints, the solution of LP relaxation must ensure that $\sum_{1\le j\le |h_i|} y^L_{p(i,j)} \ge |h_i| - 1$ $\forall y^L_{p(i,j)} \in [0,1]$. In an extreme case, values from the $(k+1)$st smallest variable to the greatest variable are equal to 1, and we need to distribute the value $k-1$ into k smaller variables. We know that the value of the largest element in the k smaller variables is at least $(k-1)/k$. ∎

4. 1.5(opt+w_{max}) approximation. In this section, we apply a $(2/3)$-rounding approach to generate an approximate solution $\varphi^R = \Phi(Y^R)$, where we assign $y^R_{p(i,j)} = 1$ if $y^L_{p(i,j)} \ge 2/3$, and $y^R_{p(i,j)} = 0$ otherwise. Lemma 4.1 states that the maximum congestion of the rounding approach is at most 1.5 times the optimal congestion of the WHEC problem when the value of the smallest LP variable in $\{y^L_{p(i,j)} \mid 1 \le j \le |h_i|\}$ is at most $1/3$; i.e., $\min_{1\le j\le |h_i|}\{y^L_{p(i,j)}\} \le 1/3$ for every hyperedge $h_i \in E_h$, $i \in \{1,2,\ldots,m\}$.

LEMMA 4.1. *The maximum congestion of the $(2/3)$-rounding approach is at most 1.5 times the optimum if $\min_{1\le j\le |h_i|}\{y^L_{p(i,j)}\} \le 1/3$ for every hyperedge $h_i \in E_h$, $i \in \{1,2,\ldots,m\}$.*

Proof. Let $\varphi^R = \Phi(Y^R)$ be the solution of the $(2/3)$-rounding approach. First, we show that $\Phi(Y^R)$ is a feasible solution if the condition of $\min_{1\le j\le |h_i|}\{y^L_{p(i,j)}\} \le 1/3$

$\forall h_i \in E_h$ is satisfied. According to connectivity constraints, any solution to LP relaxation must satisfy that $\sum_{1\le j\le|h_i|} y^L_{p(i,j)} \ge |h_i| - 1$ for every hyperedge $h_i \in E_h$, $i \in \{1, 2, \ldots, m\}$. If the value of the smallest LP variable in $\{y^L_{p(i,j)} \mid 1 \le j \le |h_i|\}$ is at most 1/3, we know that the value of the second smallest LP variable is at least 2/3. This implies that the sum of rounded values $\sum_{1\le j\le|h_i|} y^R_{p(i,j)}$ is at least $|h_i| - 1$ $\forall i \in \{1, 2, \ldots, m\}$. Each hyperedge $h_i \in E_h$ is continuously connected, and hence the solution $\Phi(Y^R)$ is feasible. Next, we show that $\varphi^R \le 1.5\varphi^*$. If $y^R_{p(i,j)}$ is rounded to 1, then $y^L_{p(i,j)}$ has a value at least 2/3. Therefore, we have $y^R_{p(i,j)} \le (3/2)y^L_{p(i,j)}$ $\forall i, j$. From the capacity constraints, we have $\sum_{p(i,j)\in P(e)} y^R_{p(i,j)} w_i \le \varphi^R \le \sum_{p(i,j)\in P(e)} (3/2) y^L_{p(i,j)} w_i \le (3/2)\varphi^L \le (3/2)\varphi^*$ $\forall e \in E_c$. □

When $\min_{1\le j\le|h_i|}\{y^L_{p(i,j)}\} \le 1/3$ $\forall h_i \in E_h$, the (2/3)-rounding approach generates a feasible solution whose value is at most 1.5 times the optimum. However, the solution of the (2/3)-rounding algorithm may be infeasible if the condition is unsatisfied. In this case, we will propose a $1.5(opt + w_{max})$ approximation algorithm on the condition that $\min_{1\le j\le|h_i|}\{y^L_{p(i,j)}\} > 1/3$, $\exists h_i \in E_h$.

From Lemma 3.1, the value of the kth smallest variable in $\{y^L_{p(i,j)} \mid 1 \le j \le |h_i|\}$ is at least $(k-1)/k$ for each hyperedge $h_i \in E_h$. If the (2/3)-rounding algorithm is applied to generate an approximate solution, the third smallest variable and these $|h_i|-3$ larger variables will certainly be rounded to 1. If the condition $\min_{1\le j\le|h_i|}\{y^L_{p(i,j)}\} \le 1/3$ is unsatisfied, values of the first and second smallest variables are both less than 2/3 and will be truncated to 0 by the (2/3)-rounding algorithm. In this case, the hyperedge h_i is disconnected, and we need to round either the first or second smallest variable to 1 for ensuring the connectivity of the hyperedge.

Let D be the set of disconnected hyperedges in the output of the (2/3)-rounding algorithm, i.e., $D = \{h_i | \sum_{1\le j\le|h_i|} y^R_{p(i,j)} = |h_i| - 2, 1 \le i \le m\}$. The problem of rounding selection for the smallest two variables on $h_i \in D$ can be translated into the ring loading problem without demand splitting. The ring loading problem was studied by Cosares and Saniee [11], Schrijver, Seymour, and Winkler [12], Wilfong and Winkler [14], Khanna [13], and Myung [15]. When each demand must be entirely routed in either of the two directions, the ring loading problem is NP-complete. Here we need only to consider the unsplitting subproblem of the ring loading problem. In this case, each separated variable in D must be merged and entirely routed in either of the two directions.

Let $y^L_{g(i,k)}$ be the kth smallest LP variable for the hyperedge h_i, where $g(i,k)$ represents the corresponding adjacent path for the kth smallest variable. If the adjacent path $g(i,2)$ for the second smallest variable is connected around the cycle from vertex s_i to vertex t_i in the clockwise direction, we rearrange the adjacent path $g(i,1)$ for connecting in the other way around the cycle from t_i to s_i. Now we define new variables as $y^M_{g(i,2)} = \frac{3}{2} y^L_{g(i,2)}$ and $y^M_{g(i,1)} = 1 - y^M_{g(i,2)}$. For each disconnected hyperedge $h_i \in D$, we need to round either $y^M_{g(i,1)}$ or $y^M_{g(i,2)}$ to 1 for ensuring the connectivity of the hyperedge h_i. Our objective is to find an optimal rounding assignment of $y^R_{g(i,1)}$ and $y^R_{g(i,2)}$ for each disconnected hyperedge $h_i \in D$ such that the maximum increment of weighted load, denoted as Δ, over any physical link in the cycle is minimized. The next lemma shows that the translation is valid.

LEMMA 4.2. *For each disconnected hyperedge $h_i \in D$, we have $(y^M_{g(i,2)} + y^R_{g(i,2)})w_i \le \frac{3}{2} y^L_{g(i,2)} w_i$ and $(y^M_{g(i,1)} + y^R_{g(i,j)})w_i \le \frac{3}{2} y^L_{g(i,j)} w_i$ $\forall j \ne 2$.*

Proof. Consider a disconnected hyperedge $h_i \in D$ from the output of (2/3)-

rounding algorithm. From Lemma 3.1, the value of the third smallest LP variable $y^L_{g(i,3)}$ of h_i is at least 2/3. Since the hyperedge h_i is disconnected, both the first and second smallest LP variables should be less than 2/3, i.e., $y^L_{g(i,1)} \le y^L_{g(i,2)} < \frac{2}{3}$. Therefore, we have $y^R_{g(i,1)} = 0$, $y^R_{g(i,2)} = 0$, and $y^R_{g(i,j)} = 1\ \forall j \ge 3$. It is trivial that $(y^M_{g(i,2)} + y^R_{g(i,2)})w_i \le \frac{3}{2}y^L_{g(i,2)}w_i$, since we define $y^M_{g(i,2)} = \frac{3}{2}y^L_{g(i,2)}$ and we know that $y^R_{g(i,2)} = 0$. Next, we show that $(y^M_{g(i,1)} + y^R_{g(i,1)})w_i \le \frac{3}{2}y^L_{g(i,1)}w_i$. From the definition, we have $y^M_{g(i,1)} = 1 - \frac{3}{2}y^L_{g(i,2)}$. Since $y^R_{g(i,1)} = 0$ and we know that $y^L_{g(i,1)} + y^L_{g(i,2)} \ge 1$ from Lemma 3.1, we have $(y^M_{g(i,1)}+y^R_{g(i,1)})w_i = (1-\frac{3}{2}y^L_{g(i,2)})w_i \le (1-\frac{3}{2}(1-y^L_{g(i,1)}))w_i < \frac{3}{2}y^L_{g(i,1)}w_i$. Finally, we show that $(y^M_{g(i,1)} + y^R_{g(i,j)})w_i \le \frac{3}{2}y^L_{g(i,j)}w_i$, $3 \le j \le |h_i|$. Since $y^R_{g(i,1)} = 0$, $y^L_{g(i,1)} < \frac{2}{3}$ and we know that $y^L_{g(i,1)} + y^L_{g(i,2)} + y^L_{g(i,3)} \ge 2$ from Lemma 3.1, we have $y^L_{g(i,2)} \ge 2 - y^L_{g(i,1)} - y^L_{g(i,3)} > 2 - \frac{2}{3} - y^L_{g(i,3)} = \frac{4}{3} - y^L_{g(i,3)}$. Then, for $j \ge 3$, we have $y^L_{g(i,2)} \ge \frac{4}{3} - y^L_{g(i,j)}$ since $y^L_{g(i,j)} \ge y^L_{g(i,3)}$. We thus have $(y^M_{g(i,1)} + y^R_{g(i,j)})w_i = (1 - \frac{3}{2}y^L_{g(i,2)} + y^R_{g(i,j)})w_i \le (2 - \frac{3}{2}(\frac{4}{3} - y^L_{g(i,j)}))w_i = \frac{3}{2}y^L_{g(i,j)}w_i$ for any $j \ge 3$. □

Next, we extend the merging and sequential rounding techniques [12, 14] from the ring loading problem to ensure $\sum_{p(i,j)\in P(e)} y^R_{p(i,j)}w_i \le \frac{3}{2}(w_{max}+\sum_{p(i,j)\in P(e)} y^L_{p(i,j)}w_i) \le \frac{3}{2}(opt + w_{max})\ \forall e \in E_c$. Here *opt* represents the optimal value of the problem and w_{max} denotes the largest weight of hyperedges, i.e., $opt = \varphi^*$ and $w_{max} = \max\{w_i \mid 1 \le i \le m\}$. In order to apply the merging and sequential rounding techniques for the unsplitting problem, we represent the notation of adjacent paths $g(i,1)$ and $g(i,2)$ for each disconnected hyperedge $h_i \in D$ in clockwise order around the cycle. Let $s(i,1)$ and $s(i,2)$, respectively, be the first and second adjacent paths of h_i in clockwise order around the cycle (sequencing from vertex 1). Now, given the adjacent paths $s(i,1)$ and $s(i,2)$ with translative variables $y^M_{s(i,1)}$ and $y^M_{s(i,2)}$ for all disconnected hyperedge $h_i \in D$, our objective is to find an optimal rounding assignment of $y^R_{s(i,1)}$ and $y^R_{s(i,2)}$ such that the maximum increment of weighted load over any physical link in the cycle is minimized. Two disconnected hyperedges h_i and h_j are said to be *crossing* if both their two adjacent paths $s(i,1)$, $s(i,2)$ and $s(j,1)$, $s(j,2)$ intersect on at least one physical link. Otherwise, they are said to be *noncrossing*. The next lemma describes how to merge two noncrossing hyperedges into a connected hyperedge without increasing the total load over any physical link.

LEMMA 4.3. *If two disconnected hyperedges are noncrossing, they can be merged into a connected hyperedge without increasing the total load over any physical link.*

Proof. Suppose that two disconnected hyperedges h_i and h_j are noncrossing. Without loss of generality, we assume that their noncrossing adjacent paths are given as Figure 2(a). In this figure, we assume that $a = w_i$, $a_1 = y^M_{s(i,1)}w_i$, $a_2 = y^M_{s(i,2)}w_i$, $b = w_j$, $b_1 = y^M_{s(j,1)}w_j$, and $b_2 = y^M_{s(j,2)}w_j$, respectively. If $y^M_{s(j,2)}w_j \ge (1 - y^M_{s(i,1)})w_i$ (see Figure 2(b)), we output an unsplitting assignment to ensure the connectivity of the hyperedge h_i by setting $y^R_{s(i,1)} = y^M_{s(i,1)} = 1$, $y^R_{s(i,2)} = y^M_{s(i,2)} = 0$, and $D = D - \{h_i\}$. Then we define a new splitting assignment for the hyperedge h_j by setting $y^M_{s(j,1)} = y^M_{s(j,1)} + y^M_{s(i,2)}(w_i/w_j)$ and $y^M_{s(j,2)} = y^M_{s(j,2)} - (1 - y^M_{s(i,1)})(w_i/w_j)$. The new load on each physical link is either unchanged or reduced. And, the new splitting assignment for the hyperedge h_j keeps the same property of $y^M_{s(j,1)} + y^M_{s(j,2)} = 1$. On the other hand, if $y^M_{s(j,2)}w_j < (1-y^M_{s(i,1)})w_i$ (see Figure 2(c)), we output an unsplitting assignment to ensure the connectivity of hyperedge h_j by setting $y^R_{s(j,1)} = y^M_{s(j,1)} = 1$,

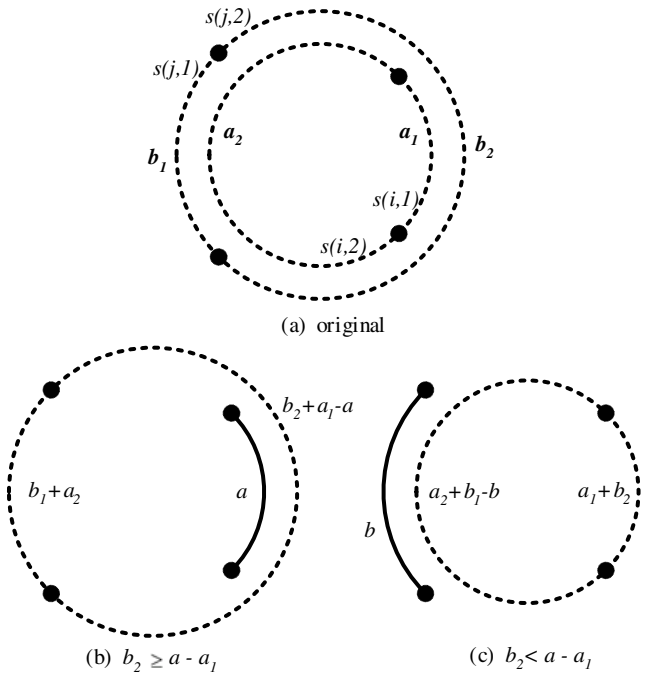


FIG. 2. *Merge two disconnected hyperedges into a connected hyperedge.*

$y^R_{s(j,2)} = y^M_{s(j,2)} = 0$, and $D = D - \{h_j\}$. Similarly, we define a new splitting assignment for the hyperedge h_i by setting $y^M_{s(i,1)} = y^M_{s(i,1)} + y^M_{s(j,2)}(w_j/w_i)$ and $y^M_{s(i,2)} = y^M_{s(i,2)} - (1 - y^M_{s(j,1)})(w_j/w_i)$. In this case, the load on each physical link and the property for the hyperedge h_i are also properly maintained. □

The merging procedure can be performed repeatedly without increasing the maximum load over any physical link; i.e., the maximum load increment Δ in the cycle is zero. Since each time we reduce by one the total number of disconnected hyperedges, this procedure will terminate in at most $|D|$ steps until no two disconnected hyperedges contain a pair of noncrossing adjacent paths. Nevertheless, not all adjacent paths of disconnected hyperedges are noncrossing. Let S be the set of remnant disconnected hyperedges. We perform a sequential rounding technique to round each remnant variable in the clockwise direction around the cycle. In the following algorithm, we combine the (2/3)-rounding, merging, and sequential rounding techniques for ensuring an approximation bound of $1.5(opt + w_{max})$. Here *opt* represents the optimal value of the problem and w_{max} denotes the largest weight of hyperedges, i.e., $w_{max} = \max\{w_i \mid 1 \le i \le m\}$.

ALGORITHM 1. *The improved* (2/3)*-rounding algorithm.*

Step (1). *Solve optimally the LP relaxation of the MILP formulation. Let the optimal solution be* $\Phi(Y^L)$, *where* $Y^L = [y^L_{p(i,j)}]$ *and* $0 \le y^L_{p(i,j)} \le 1$.

Step (2). $\forall i, j, let y^R_{p(i,j)} = 1$ *if* $y^L_{p(i,j)} \ge 2/3$, *and* $y^R_{p(i,j)} = 0$ *otherwise.*

Step (3). *Let* $D = \{h_i \mid \sum_{1 \le j \le |h_i|} y^R_{p(i,j)} = |h_i| - 2, 1 \le i \le m\}$ *be the set of disconnected hyperedges. Let* $g(i,j)$ *denote the adjacent path with the* j*th smallest LP variable for the hyperedge* $h_i \in D$. *Rearrange the adjacent path* $g(i,1)$ *and assign* $y^M_{g(i,2)} = \frac{3}{2} y^L_{g(i,2)}$ *and* $y^M_{g(i,1)} = 1 - y^M_{g(i,2)}$, *respectively.*

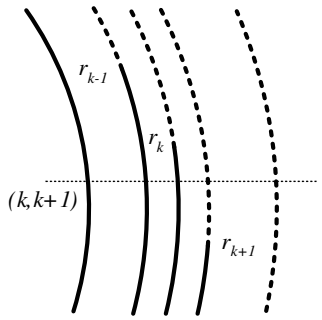


FIG. 3. *The link* $(k, k+1)$ *is crossed by the adjacent paths* $s(r_1, 1)$, $s(r_2, 1), \ldots, s(r_k, 1)$, $s(r_{k+1}, 2)$, $s(r_{k+2}, 2), \ldots, s(r_{|S|}, 2)$.

Step (4). *Given the adjacent paths* $g(i, j)$ *and translative variables* $y^M_{g(i,j)}$ $\forall h_i \in D$, $j \in \{1, 2\}$, *perform the merging procedure on* D *repeatedly until only crossing adjacent paths remain.*

Step (5). *Let* S *be the set of remnant hyperedges. For each hyperedge* $h_i \in S$, *let* $y^M_{s(i,1)}$ *and* $y^M_{s(i,2)}$, *respectively, be the first and second variables of* h_i *in clockwise order. And let* $r_1, r_2, \ldots, r_{|S|}$ *be the labels of hyperedges in* S *in clockwise sequential order.*

Step (6). *Let* $y^R_{s(r_1,1)} = 1$ *and* $y^R_{s(r_1,2)} = 0$ *if* $y^M_{s(r_1,1)} > 0.5$; *otherwise let* $y^R_{s(r_1,1)} = 0$ *and* $y^R_{s(r_1,2)} = 1$.

Step (7). $\forall 2 \le k \le |S|$, *if* $(1-y^M_{s(r_k,1)})w_k + \sum_{i\in\{r_1,\ldots,r_{k-1}\}}(y^R_{s(i,1)} - y^M_{s(i,1)})w_i < w_{max}/2$, *then let* $y^R_{s(r_k,1)} = 1$ *and* $y^R_{s(r_k,2)} = 0$; *otherwise let* $y^R_{s(r_k,1)} = 0$ *and* $y^R_{s(r_k,2)} = 1$. *Due to the complementary nature, we have* $\sum_{i\in\{r_1,\ldots,r_k\}}(y^R_{s(i,1)} - y^M_{s(i,1)})w_i \in [-\frac{w_{max}}{2}, \frac{w_{max}}{2})$ *and* $\sum_{i\in\{r_1,\ldots,r_k\}}(y^R_{s(i,2)} - y^M_{s(i,2)})w_i \in (-\frac{w_{max}}{2}, \frac{w_{max}}{2}]$.

Step (8). *Output the approximate solution* $\varphi^R = \Phi(Y^R)$ *of the maximum congestion.*

THEOREM 4.4. *The improved* (2/3)*-rounding algorithm has an approximation bound of* $1.5(opt + w_{max})$, *where opt represents the optimal value of the problem and* w_{max} *denotes the largest weight of hyperedges.*

Proof. Step (6) ensures that $(y^R_{s(r_1,1)} - y^M_{s(r_1,1)})w_{r_1} \in [-\frac{w_{r_1}}{2}, \frac{w_{r_1}}{2})$ and $(y^R_{s(r_1,2)} - y^M_{s(r_1,2)})w_{r_1} \in (-\frac{w_{r_1}}{2}, \frac{w_{r_1}}{2}]$. By induction, if the partial sum $\sum_{i\in\{r_1,\ldots,r_{k-1}\}}(y^R_{s(i,1)} - y^M_{s(i,1)})w_i$ from r_1 to r_{k-1} is given and places in the interval $[-\frac{w_{max}}{2}, \frac{w_{max}}{2})$, then we define $y^R_{s(k,1)}$ sequentially around the cycle from r_1 to $r_{|S|}$ by Step (7) for ensuring $\sum_{i\in\{r_1,\ldots,r_k\}}(y^R_{s(i,1)} - y^M_{s(i,1)})w_i \in [-\frac{w_{max}}{2}, \frac{w_{max}}{2})$. Due to the complementary nature, we also have $\sum_{i\in\{r_1,\ldots,r_k\}}(y^R_{s(i,2)} - y^M_{s(i,2)})w_i \in (-\frac{w_{max}}{2}, \frac{w_{max}}{2}]$.

For any physical link $(k, k+1) \in E_c$, we assume without loss of generality that it is crossed by two partitions of sequential hyperedges $r_1, r_2, \ldots, r_k$ and $r_{k+1}, r_{k+2}, \ldots, r_{|S|}$, as shown in Figure 3. Therefore, the increment of congestion of the link $(k, k+1)$ is at most $\Delta_k = \sum_{i\in\{r_1,\ldots,r_k\}}(y^R_{s(i,1)} - y^M_{s(i,1)})w_i + \sum_{i\in\{r_{k+1},\ldots,r_{|S|}\}}(y^R_{s(i,2)} - y^M_{s(i,2)})w_i = \sum_{i\in\{r_1,\ldots,r_k\}}(y^R_{s(i,1)} - y^M_{s(i,1)})w_i + \sum_{i\in\{r_1,\ldots,r_{|S|}\}}(y^R_{s(i,2)} - y^M_{s(i,2)})w_i - \sum_{i\in\{r_1,\ldots,r_k\}}(y^R_{s(i,2)} - y^M_{s(i,2)})w_i \le \frac{w_{max}}{2} + \frac{w_{max}}{2} - (-\frac{w_{max}}{2}) = \frac{3}{2}w_{max}$. We thus have $\Delta = \max_k\{\Delta_k\} \le \frac{3}{2}w_{max}$. Therefore, from Lemma 4.1, we have an approximation bound of $1.5(opt + w_{max})$. ∎

5. (1.5+ε) approximation. The improved (2/3)-rounding algorithm generates a feasible solution whose value is at most $1.5(opt+w_{max})$. Clearly, if the optimal value *opt* of maximum congestion is large, the additional term of the maximum weight w_{max} of hyperedges can be translated into a relatively small deviation in value from the optimal solution. However, the value of optimal solution may be near the maximum weight of hyperedges, and the additive error term could be a significant deviation from the optimum. In this section we will apply the PTAS in [13] to show that for any fixed $\varepsilon > 0$, there is a polynomial time algorithm that yields a $(1.5 + \varepsilon)$ approximation.

Let φ^T denote the value of approximate solution of the (1/2)-rounding algorithm [1]. The approximate solution of the (1/2)-rounding algorithm is always feasible, and it ensures that $\varphi^T \leq 2\varphi^*$. On the other hand, if the improved (2/3)-rounding algorithm is applied to generate an approximate solution, we need to round either the first or second smallest variable of a disconnected hyperedge to 1 for ensuring the connectivity. In Algorithm 1, we denote S as the set of these disconnected hyperedges. A hyperedge in S is said to be *heavy* if its weight is at least $\frac{1}{3}\varepsilon\varphi^T$ and *light* otherwise. Let S_H and S_L, respectively, denote the subsets of heavy and light hyperedges in S. The following lemma bounds the total number of heavy hyperedges in S_H.

LEMMA 5.1. *The set S_H contains at most $\frac{3n}{2\varepsilon}$ hyperedges, i.e., $|S_H| \leq \frac{3n}{2\varepsilon}$.*

Proof. Each hyperedge in S_H has weight at least $\frac{1}{3}\varepsilon\varphi^T \leq \frac{2}{3}\varepsilon\varphi^*$. In an ideal case, these hyperedges are routed in one hop and shared out equally among n physical links in the cycle. Obviously, the maximum congestion on any physical link is at least $\frac{2}{3}\varepsilon\varphi^* \frac{|S_H|}{n}$. This gives a lower bound on the maximum congestion for embedding all hyperedges in S_H. Hence we have $\frac{2}{3}\varepsilon\varphi^* \frac{|S_H|}{n} \leq \varphi^*$. Rearranging the term, we get $|S_H| \leq \frac{3n}{2\varepsilon}$. ∎

For every disconnected hyperedge in S_H, we need to choose one of two adjacent paths for ensuring the connectivity. If the longer of the two adjacent paths is selected for connecting a disconnected hyperedge along the cycle, we say that the hyperedge is routed in the long way. The next lemma bounds the total number of heavy hyperedges that could be routed in the long way for any optimal routing.

LEMMA 5.2. *In any optimal routing, at most $3/\varepsilon$ heavy hyperedges are embedded in the long way.*

Proof. Suppose to the contrary that more than $3/\varepsilon$ heavy hyperedges were routed in the long way in the optimal solution. Since the hyperedge must be routed at least $\lceil n/2 \rceil$ physical links in the long way, the total weight induced by those heavy hyperedges in the long way is more than $\lceil n/2 \rceil \times \frac{2}{3}\varepsilon\varphi^* \times 3/\varepsilon \geq n\varphi^*$. By the pigeonhole principle, there must be some links with congestion more than the optimal value $\frac{n\varphi^*}{n} = \varphi^*$, which contradicts the optimality of the optimal routing. ∎

Since the largest weight of hyperedges in S_L is less than $\frac{1}{3}\varepsilon\varphi^T \leq \frac{2}{3}\varepsilon\varphi^*$, we can perform the clockwise sequential rounding algorithm on S_L to generate an approximation with a bound of $\frac{3}{2}w_{max} < \frac{3}{2} \times \frac{2}{3}\varepsilon\varphi^* = \varepsilon\varphi^*$. On the other hand, we need to exhaustively try all possible ways to find the optimal solution for embedding all hyperedges in S_H. Fortunately, the time complexity of the exhaustive search is bounded by a polynomial function of the problem size.

LEMMA 5.3. *For an exhaustive search on S_H, the total number of times for searching is bounded by $k(\frac{en}{2})^k$, where $k = 3/\varepsilon$.*

Proof. Since the set S_H contains at most $\frac{3n}{2\varepsilon}$ hyperedges, we need only to choose at most $3/\varepsilon$ heavy hyperedges from S_H for embedding them in the long way. Let $k = 3/\varepsilon$. The number of choices is bounded by $\sum_{i=0}^{k} \binom{\frac{kn}{2}}{i}$. From the well-known

inequality $\binom{x}{y} \leq (\frac{ex}{y})^y$, we have

$$\sum_{i=0}^{k} \binom{\frac{kn}{2}}{i} \leq \sum_{i=0}^{k} \left(\frac{e \times kn}{2i} \right)^i \leq k \left(\frac{en}{2} \right)^k.$$

Thus, the total running time of exhaustive search on S_H is bounded by a polynomial function of n. □

THEOREM 5.4. *For any fixed $\varepsilon > 0$, there is a polynomial time algorithm whose approximate solution is within $(1.5 + \varepsilon)$ times the optimum for the general WHEC problem.*

6. Conclusion. This paper concerns the problem of WHEC and focuses on the impact of the maximum congestion. We formulate the problem as a MILP and propose an improved (2/3)-rounding algorithm to provide a solution with an approximation bound of $1.5(opt + w_{max})$, where opt represents the optimal value of the WHEC problem and w_{max} denotes the largest weight of hyperedges. If the value of optimal solution is near the maximum weight of hyperedges, then the additive error term w_{max} could be a significant deviation from the optimum. For any fixed $\varepsilon > 0$, we also present a polynomial time approximation algorithm to provide an embedding whose congestion is at most $(1.5 + \varepsilon)$ times the optimum. To our knowledge, this is the best approximation bound known for the general problem of embedding weighted hypergraph in a cycle. However, the main problem with the LP-based approximation is the time required to solve the linear program. To improve the efficiency, it is worthwhile to look into ways of making better use of the congestion information for obtaining a simple heuristic algorithm with a tighter approximation bound.

Acknowledgments. The author thanks the editor and anonymous reviewers for their constructive comments.

REFERENCES

[1] S. L. LEE AND H.-J. HO , *On minimizing the maximum congestion for weighted hypergraph embedding in a cycle*, Inform. Process. Lett., 87 (2003), pp. 271–275.
[2] T. GONZALEZ AND S. L. LEE, *A linear time algorithm for optimal wiring around a rectangle*, J. Assoc. Comput. Mach., 35 (1998), pp. 810–832.
[3] J. L. GANLEY AND J. P. COHOON, *A provably good moat routing algorithm*, in Proceedings of the 6th Great Lakes Symposium on VLSI, IEEE Computer Society, Washington, DC, 1996, pp. 86–91.
[4] A. FRANK, T. NISHIZEKI, N SAITO, H. SUZUKI, AND E. TARDOS, *Algorithms for routing around a rectangle*, Discrete Appl. Math., 40 (1992), pp. 363–378.
[5] J. L. GANLEY AND J. P. COHOON, *Minimum-congestion hypergraph embedding in a cycle*, IEEE Trans. Comput., 46 (1997), pp. 600–602.
[6] T. GONZALEZ, *Improved approximation algorithms for embedding hyperedges in a cycle*, Inform. Process. Lett., 67 (1998), pp. 267–271.
[7] T. CARPENTER, S. COSARES, J. L. GANLEY, AND I. SANIEE, *A simple approximation algorithm for two problems in circuit design*, IEEE Trans. Comput., 47 (1998), pp. 1310–1312.
[8] Q. P. GU AND Y. WANG, *Efficient algorithms for minimum congestion hypergraph embedding in a cycle*, IEEE Trans. Parallel Distrib. Systems, 17 (2006), pp. 205–214.
[9] S. L. LEE AND H.-J. HO, *A 1.5 approximation algorithm for embedding hyperedges in a cycle*, IEEE Trans. Parallel Distrib. Systems, 16 (2005), pp. 481–488.
[10] X. DENG AND G. LI, *A PTAS for embedding hypergraph in a cycle (extended abstract)*, in Proceedings of the 31st International Colloquium on Automata, Languages and Programming (Turku, Finland, 2004), Springer, Berlin, 2004, pp. 433–444.
[11] S. COSARES AND I. SANIEE, *An optimization problem related to balancing loads on SONET rings*, Telecommunications Systems, 3 (1992), pp. 165–181.

[12] A. Schrijver, P. Seymour, and P. Winkler, *The ring loading problem*, SIAM J. Discrete Math., 11 (1998), pp. 1–14.
[13] S. Khanna, *A polynomial time approximation scheme for SONET ring loading problem*, Bell Labs Tech. J., 2 (1997), pp. 36–41.
[14] G. Wilfong and P. Winkler, *Ring routing and wavelength translation*, in Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, 1998, pp. 333–341.
[15] Y.-S. Myung, *An efficient algorithm for the ring loading problem with integer demand splitting*, SIAM J. Discrete Math., 14 (2001), pp. 291–298.
[16] S. Vempala and B. Vöcking, *Approximating multicast congestion*, in Proceedings of the 10th International Symposium on Algorithms and Computation, Springer, Berlin, 1999, pp. 367–372.
[17] R. Carr and S. Vempala, *Randomized metarounding*, in Proceedings of the 32nd ACM Symposium on Theory of Computing, 2000, pp. 58–62.
[18] K. Jansen and H. Zhang, *An approximation algorithm for the multicast congestion problem via minimum Steiner trees*, in Proceedings of the 3rd International Workshop on Approximation and Randomized Algorithms in Communication Networks, Carleton Scientific, Waterloo, ON, Canada, 2002, pp. 1–15.

SIAM J. OPTIM.
Vol. 18, No. 4, pp. 1501–1503

ERRATUM: MESH ADAPTIVE DIRECT SEARCH ALGORITHMS FOR CONSTRAINED OPTIMIZATION*

CHARLES AUDET†, A. L. CUSTÓDIO‡, AND J. E. DENNIS, JR.§

Abstract. In [*SIAM J. Optim.*, 17 (2006), pp. 188–217] Audet and Dennis proposed the class of *mesh adaptive direct search* (MADS) *algorithms* for minimization of a nonsmooth function under general nonsmooth constraints. The notation used in the paper evolved since the preliminary versions, and, unfortunately, even though the statement of Proposition 4.2 is correct, it is not compatible with the final notation. The purpose of this note is to show that the proposition is valid.

Key words. mesh adaptive direct search algorithms, constrained optimization, nonsmooth optimization

AMS subject classifications. 90C30, 90C56, 65K05, 49J52

DOI. 10.1137/060671267

In [1] Audet and Dennis proposed the class of *mesh adaptive direct search* (MADS) *algorithms* for minimization of a nonsmooth function under general nonsmooth constraints. The paper contains a convergence analysis for this class of methods and proposes two variants of an implementable instance called LTMADS.

The proof that LTMADS is indeed an instance of MADS is not compatible with the notation used in the rest of the paper. We restate the proposition and propose a consistent proof.

PROPOSITION 0.1 (Proposition 4.2 of [1]). *At each iteration k, the procedure above yields a D_k and a MADS frame P_k such that*

$$P_k = \{x_k + \Delta_k^m d : d \in D_k\} \subset M_k,$$

where $\Delta_k^m > 0$ is the mesh size parameter, M_k is given by Definition 2.1 of [1], and D_k is a positive spanning set such that for each $d \in D_k$,

- *d can be written as a nonnegative integer combination of the directions in D: $d = Du$ for some vector $u \in \mathbb{N}^{n_D}$ that may depend on the iteration number k;*
- *the distance from the frame center x_k to a frame point $x_k + \Delta_k^m d \in P_k$ is bounded above by a constant times the poll size parameter: $\Delta_k^m \|d\|_\infty \leq \Delta_k^p \max\{\|d'\|_\infty : d' \in D\}$;*
- *limits (as defined in Coope and Price [2]) of convergent subsequences of the normalized sets $\overline{D_k} := \{\frac{d}{\|d\|_\infty} : d \in D_k\}$ are positive spanning sets.*

*Received by the editors October 2, 2006; accepted for publication (in revised form) July 1, 2007; published electronically January 16, 2008.

http://www.siam.org/journals/siopt/18-4/67126.html

†GERAD and Département de Mathématiques et de Génie Industriel, École Polytechnique de Montréal, C.P. 6079, Succ. Centre-ville, Montréal H3C 3A7, QC, Canada (Charles.Audet@gerad.ca, http://www.gerad.ca/Charles.Audet). This author's research was supported by NSERC grant 239436-01, AFOSR FA9550-04-1-0235, the Boeing Company, and ExxonMobil.

‡Departamento de Matemática, FCT-UNL, Quinta da Torre, 2829-516 Caparica, Portugal (alcustodio@fct.unl.pt, http://ferrari.dmat.fct.unl.pt/personal/alcustodio/). This author's research was supported by Centro de Matemática da Universidade de Coimbra and the FCT under grant POCI/MAT/59442/2004.

§Computational and Applied Mathematics Department, Rice University, 8419 42nd Ave. SW, Seattle, WA 98136 (dennis@rice.edu, http://www.caam.rice.edu/∼dennis). This author's research was supported by AFOSR FA9550-04-1-0235, the Boeing Company, and ExxonMobil.

Proof. In order to construct the set of directions D_k, the algorithm builds matrices at iteration k that should be called L_k, B_k, and B'_k. To ease the presentation, we omit the index k in the proof of the two first bullets. The index k reappears in the proof of the last bullet since this last result involves limits as k goes to infinity.

By the construction in [1], L is a lower triangular $(n-1)\times(n-1)$ matrix where each term on the diagonal is either plus or minus 2^ℓ, and the lower components are randomly chosen from the discrete set $\{-2^\ell+1,-2^\ell+2,\ldots,2^\ell-1\}$, with ℓ an integer that satisfies $2^\ell = 1/\sqrt{\Delta_k^m}$. The rules for updating the mesh size parameter Δ_k^m ensure that $\ell \in \mathbb{N}$. It follows that L is a basis in $\mathbb{R}^{n-1}$ with $|\det(L)| = 2^{\ell(n-1)}$. Let $\{p_1, p_2, \ldots, p_{n-1}\}$ be a random permutation of the set $\{1,2,\ldots,n\}\setminus\{\hat{\imath}\}$, where $\{\hat{\imath}\}$ is defined in [1]. The elements of the matrix B are defined as

$$\begin{array}{ll} B_{p_i,j} = L_{i,j} & \text{for } i,j = 1,2,\ldots,n-1,\\ B_{\hat{\imath},j} = 0 & \text{for } j = 1,2,\ldots,n-1,\\ B_{i,n} = b_i(\ell) & \text{for } i = 1,2,\ldots,n,\end{array}$$

where $b_i(\ell)$ is a vector that depends only on the value of the mesh size parameter and not on the iteration number (see section 4.1 of [1]). It follows that B is a permutation of the rows and the columns of a lower triangular matrix whose diagonal elements are either -2^ℓ or 2^ℓ. Therefore B is a basis in $\mathbb{R}^n$ and $|\det(B)| = 2^{\ell n}$.

The square matrix B' is obtained by permuting the columns of B, and therefore the columns of B' form a basis of $\mathbb{R}^n$. Furthermore, $|\det(B')| = |\det(B)| = 2^{\ell n}$.

One of the proposed versions of LTMADS uses a minimal positive basis at every iteration, and the other variant uses a maximal positive basis at every iteration. The columns of $[B' \;\; -b']$ with $b'_i = \sum_{j\in N} B'_{ij}$ define a minimal positive basis, and the columns of $[B' \;\; -B']$ define a maximal positive basis [3].

Therefore, if $D_k = [B' \;\; -b']$ or if $D_k = [B' \;\; -B']$, then all entries of D_k are integers in the interval $[-n2^\ell, n2^\ell]$ or in the interval $[-2^\ell, 2^\ell]$, respectively. It follows that each column d of D_k can be written as a nonnegative integer combination of the columns of $D = [I \;\; -I]$. Hence, the frame defined by D_k is on the mesh M_k.

Two cases must be considered to show the second bullet. Recall that with LTMADS, the poll size parameter Δ_k^p (see [1]) is defined differently depending on whether minimal or maximal positive bases are used. If the maximal positive basis construction is used, then $\|\Delta_k^m d\|_\infty = \Delta_k^m \|d\|_\infty = \sqrt{\Delta_k^m} = \Delta_k^p$. If the minimal positive basis construction is used, then $\|\Delta_k^m d\|_\infty = \Delta_k^m \|d\|_\infty \le n\sqrt{\Delta_k^m} = \Delta_k^p$. The proof of the second bullet follows by noticing that $\max\{\|d'\|_\infty : d' \in [I \;\; -I]\} = 1$.

To show the third bullet, we will verify that the limit of the normalized sets $\overline{D_k} := \{\frac{d}{\|d\|_\infty} : d \in D_k\}$ forms a positive basis. It suffices to show that the conditions (1a), (1b), and (C1) or (C2) of Coope and Price [2] hold.

- Conditions (1a) and (1b) ensure that the limit of any convergent subsequence of the sequence of bases $\overline{B'_k} := \{\frac{d}{\|d\|_\infty} : d \in B'_k\}$ is also a basis. Condition (1a) requires that $|\det(\overline{B'_k})|$ be bounded below by a positive constant that is independent of k. In our context, $|\det(\overline{B'_k})| = 1$ for all k, and therefore this condition is satisfied. Condition (1b) is also easily satisfied since normalized directions are used. It follows that the limit of $\overline{B'_k}$ is a basis.
- Conditions (C1) and (C2) involve the columns added to each basis B'_k to form a positive basis. In the case of the maximal bases, condition (C1) is easily satisfied. For the minimal bases, (C2) holds since all the structure constants ξ (again following the definition of Coope and Price [2]) satisfy $-1 \le \xi \le -\frac{1}{n}$.

This concludes the proof. ∎

REFERENCES

[1] C. Audet and J. E. Dennis, Jr., *Mesh adaptive direct search algorithms for constrained optimization*, SIAM J. Optim., 17 (2006), pp. 188–217.
[2] I. D. Coope and C. J. Price, *Frame based methods for unconstrained optimization*, J. Optim. Theory Appl., 107 (2000), pp. 261–274.
[3] C. Davis, *Theory of positive linear dependence*, Amer. J. Math., 76 (1954), pp. 733–746.